Introduction to Survival Analysis

Dr Cameron Hurst
cphurst@gmail.com

DAMASAC and CEU, Khon Kaen University

8th September, 2557

What I will cover....

1. Introduction
   - Issues with survival data
   - Survival distributions and some important functions
2. Bivariate analysis of survival data
   - Kaplan-Meier survival curves
   - Descriptive statistics
   - Comparing groups
3. Modelling survival data
   - Introduction to Cox PH regression
   - Example: Cox PH regression
   - Assumptions of Cox PH regression
   - Extending Cox PH regression

Time-to-event data

Often we are interested in an outcome involving time to an event. For example in clinical research:

- Time to death
- Time to remission (after treatment)
- Time to recidivism (after addiction treatment)

All of these endpoints come under the heading of *survival data* (Even though not all involve death as an endpoint).
Most survival analysis are performed in the clinical setting (Only other **common** application is to the survival of products: time to failure of a computer $\rightarrow$ pricing warranties).

Issues with survival data

Analysing survival data presents a few problems not encountered in other types of analysis:

- Data truncation and censoring
- Conditional probability: To consider the probability that a subject might die at particular point in time, we have to account for the fact that they have survived to that point i.e. The probability of subject X dying on day 30 must account for the fact that they were alive on all the preceding days.

Types of censoring and truncation

Examples of censoring and truncation

Subjects that do not experience event of interest (B,C,D,E)

Incomplete follow-up

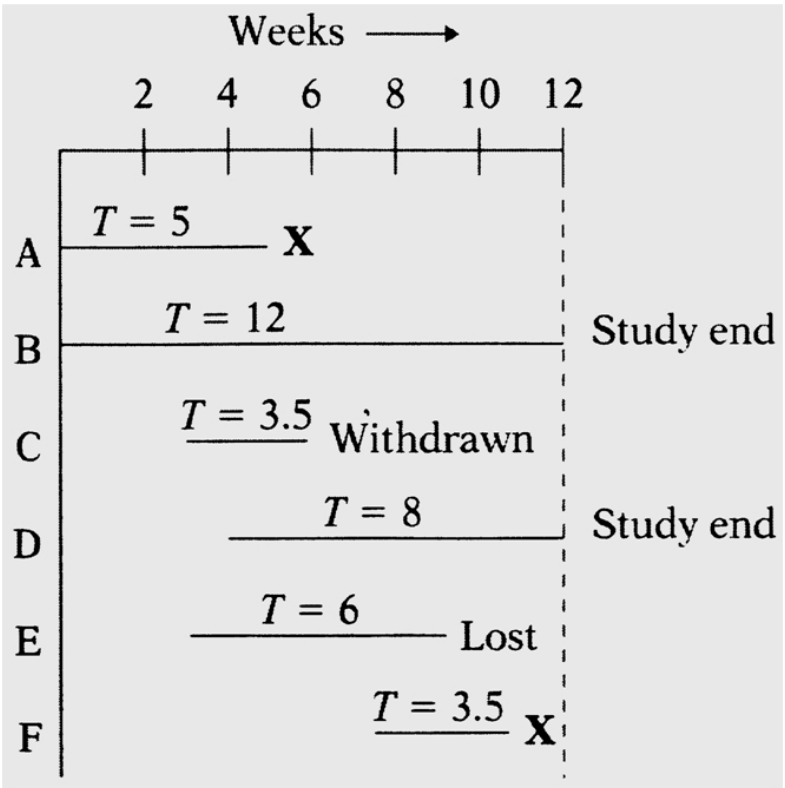
- Lost to follow-up (C)
- Withdraws from study (E)
- Dies of something other than clinical end point

Right censored (B,C,D and E)

- B and D - study ends
- C and E - withdrew/LTFU

Left truncated (A,B)

- A and B - time at risk prior to study



Distribution of survival times

First, consider the Probability Distribution (density function), $f(t)$, and Cumulative Probability Distribution, $F(t)$ of survival times, t:

$f(t)$ represents the probability of dying on a particular day, t
$F(t)$ represents the probability of any time up to a particular day, t

Example:
$f(23)$ = probability of dying day 23; and
$F(23)$ = probability of any time in the first 23 days

However, generally we are more interested in the probability that someone will **survive** up to a particular point in time, which brings us to the survival function, $S(t)$

Some important functions

Survivor function, $S(t)$ and the hazard function $h(t)$

- Rather than f(t) and F(t), focus in survival analysis is on the **survival** and **hazard** functions
- Survival function, S(t), defines the probability of surviving longer than time t. **e.g. Prob(survival time $>$ 1 yr)**

$$S(t) = P(T > t) = \int_t^\infty f(u)\,du = 1 - F(t)$$

Hazard function, $h(t)$, **instantaneous** risk of event at time t

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

Hazard function: the probability that if you survive to t, you will die in the next instant

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 (see next slide)
- 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

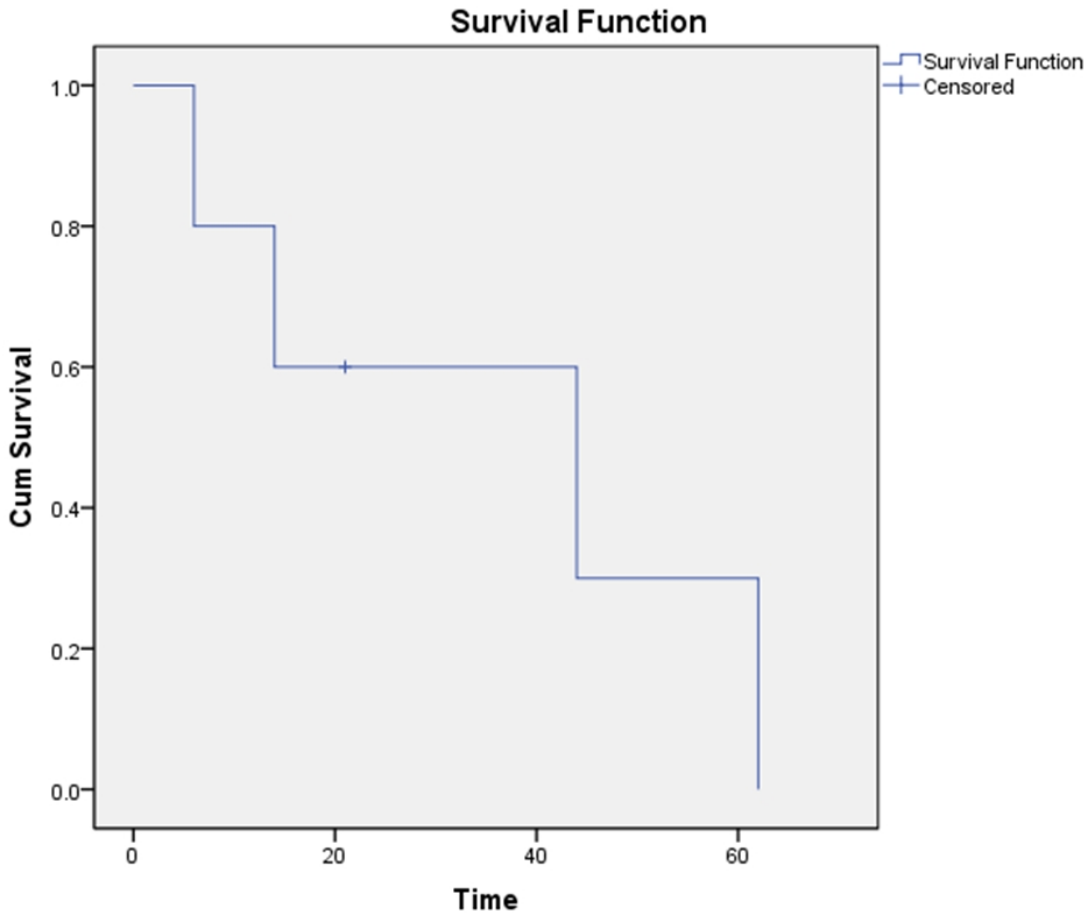
Simple example of a survival curve

Subject	*Survtime*	*Censored*
1	6	1
2	44	1
3	21	0
4	14	1
5	62	1

Subject	*calculation*	$S(t)$
$0 \leq t < 6$	$\frac{5}{5}$	1.0
$6 \leq t < 14$	$1.0 \times \frac{4}{5}$	0.8
$14 \leq t < 21$	$1.0 \times \frac{4}{5} \times \frac{3}{4}$	0.6
$21 \leq t < 44$	$1.0 \times \frac{4}{5} \times \frac{3}{4} \times \frac{3}{3}$	0.6
$44 \leq t < 62$	$1.0 \times \frac{4}{5} \times \frac{3}{4} \times \frac{3}{3} \times \frac{1}{2}$	0.3
$t > 62$	$1.0 \times \frac{4}{5} \times \frac{3}{4} \times \frac{3}{3} \times \frac{1}{2} \times \frac{0}{1}$	0

KM curve for simple example

and the resulting Kaplan-Meier curve....



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.

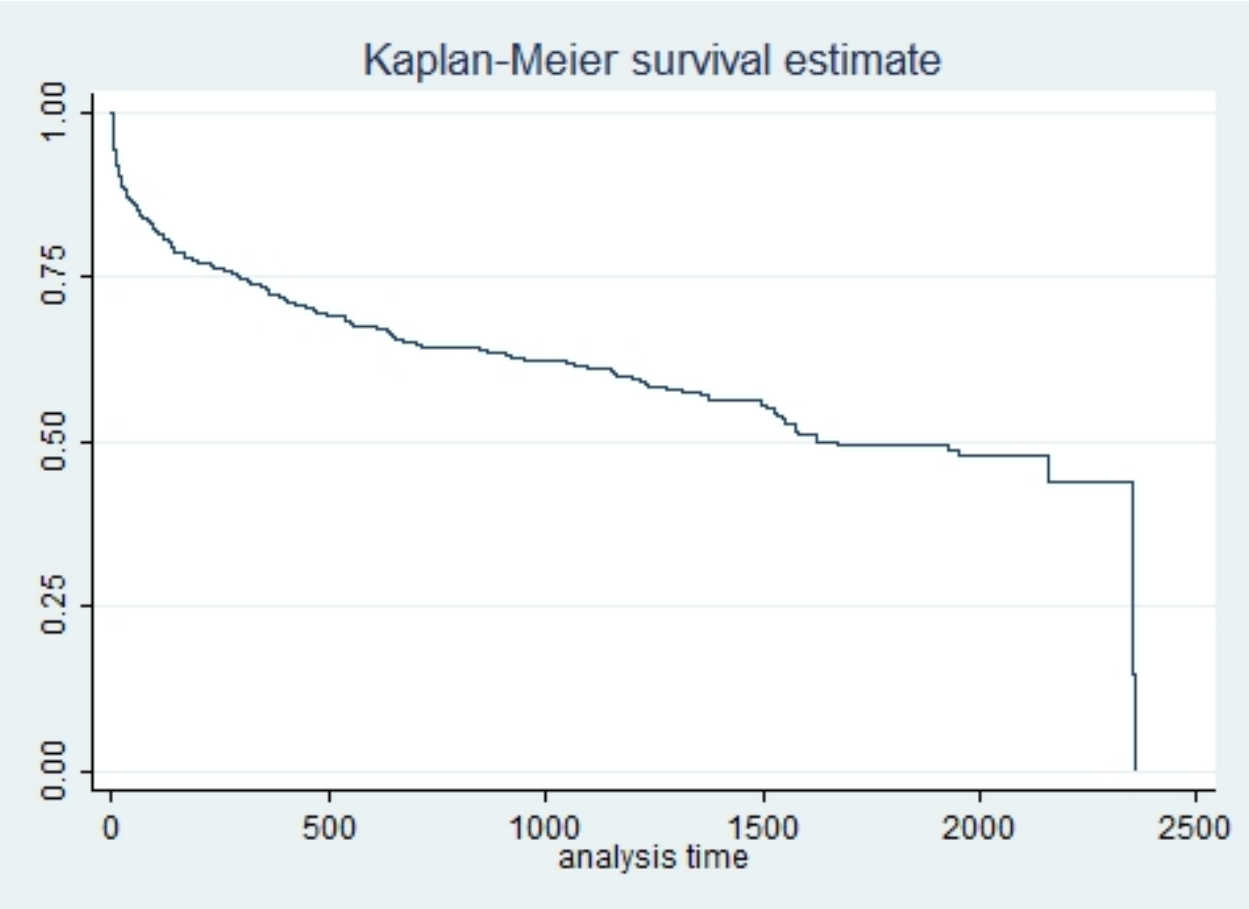
Motivating example: Worcester Heart Attack Study

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)	

KM curve for Worcester data

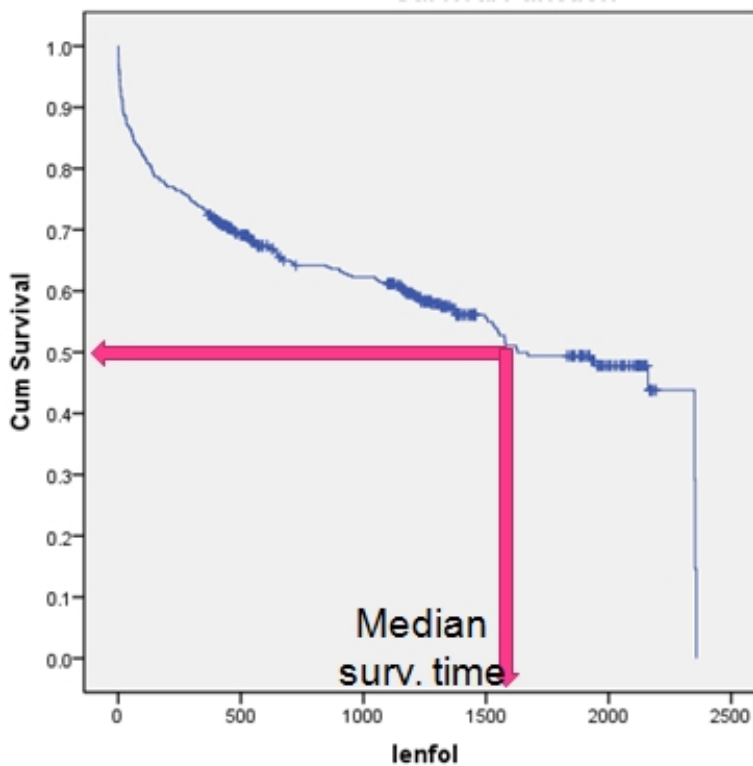
All subjects



Descriptive statistic: Median survival time

Using the KM curve

- Often clinicians want to know what is the survival time that 50% of individuals can expect to survive (Median survival time)?
- Software can calculate the median survival time, or we can just read it off the KM-curve.
- It is the time associated with the 0.5 point in cumulative survival



Descriptive statistic

Mean vs. median survival time

- As survival times are zero bound on the left (can't have survival < 0) we would expect survival times to be positively skewed.
- So, the median rather than mean survival time is the best way to represent the 'typical' survival time.

Means and Medians for Survival Time

Mean[a]				Median			
		95% Confidence Interval				95% Confidence Interval	
Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
1417.215	48.137	1322.867	1511.562	1627.000	159.555	1314.271	1939.729

a. Estimation is limited to the largest survival time if it is censored.

These values are routinely output with KM curves

Descriptive statistic

All the Quartiles

We can also obtain all the quartiles

Percentiles

25.0%		50.0%		75.0%	
Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
2353.000	79.465	1627.000	159.555	295.000	63.223

- 75% of individuals survived at least 295 days,
- 50% survived at least 1627 days (median survival time)
- 25% survived at least 2353 days

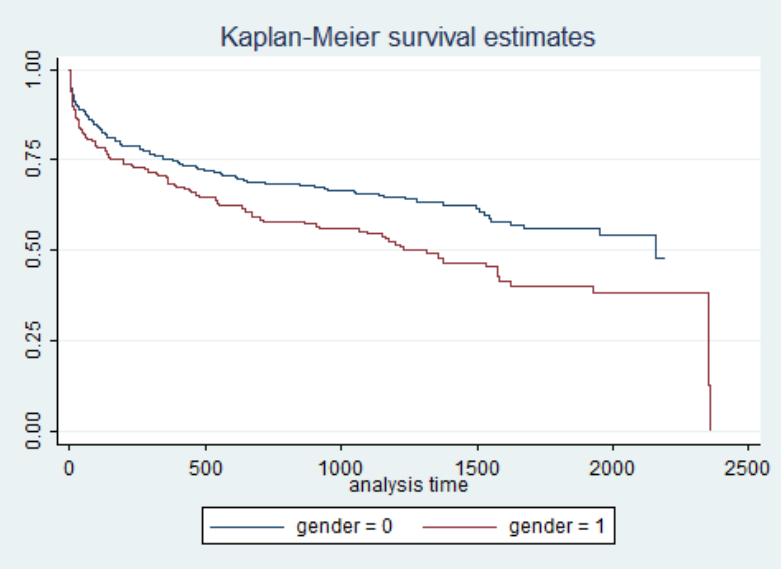
Comparing groups
Methods

- Visualizing the entire dataset well and good, but unless our sample represents a single homogeneous population, we need methods to compare groups of individuals
- A number of tests exist able to compare the survival curves arising from different groups (e.g. Survival time for those on treatment A vs. those on treatment B).
- The bivariate test we consider here is non-parametric (makes no assumption about the distribution of survival times from each group).
- This method is called the **log-rank test**.

Comparing groups: Heart attack data

Comparing KM curves

First, we can visualize group differences by generating group-specific KM curves



On the face of it, Males (gender=0) have a more favourable survival experience

Comparing groups: Heart attack data

Bivariate hypothesis testing

Now let's look at the descriptive statistics (median survival time) for each group

Percentiles

gender	25.0%		50.0%		75.0%	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
0	.	.	2160.000	.	354.000	109.695
1	2353.000	176.168	1317.000	177.039	151.000	84.201
Overall	2353.000	79.465	1627.000	159.555	295.000	63.223

- The median survival time for males (0) was 2160 days and for females (1) 1317 days.

Comparing groups: Heart attack data

Bivariate hypothesis testing

Now let's see if there is a (statistically) significant difference in the expected survival of males compared to females

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	7.791	1	.005

Test of equality of survival distributions for the different levels of gender.

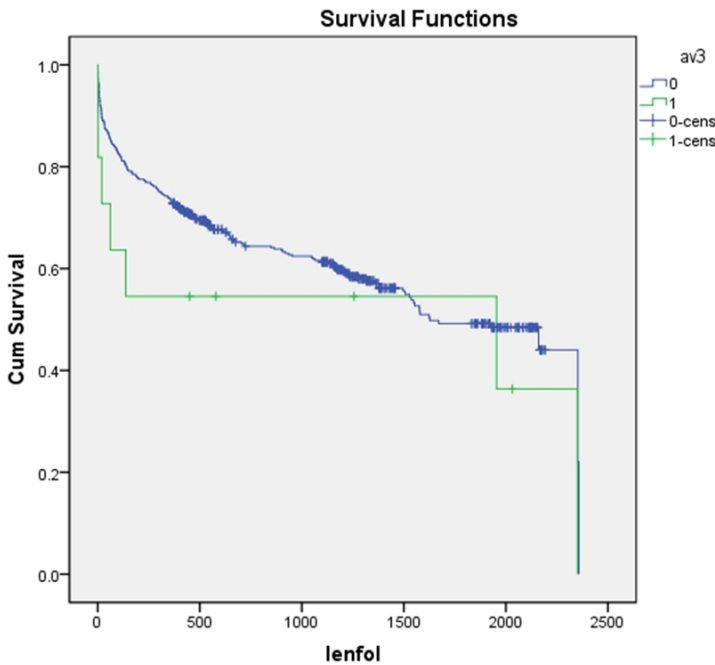
- Log-rank shows a difference in the survival experience between the groups (χ^2_{LR}=7.79, p<0.05)

Problem with Log-rank tests

Log-rank tests (and other classical bivariate tests) don't tell us much about the nature (or magnitude) of the association. We need more informative methods.

Comparing groups
A final word on KM curves

- In order to conclude a difference in the survival of two or more groups, **their survival curves should not intersect** (Regardless of Log rank test's results).
- When we compare the KM curves of those that had complete heart block (1=yes), with those that didn't (0=no), we cannot conclude a (definitive) difference in the survival experience.



Where to from here?

Modelling survival with Cox Proportional Hazards Regression

- Now we will consider true **modelling** of survival times.
- Specifically we will consider a multi-variable model that can deal with both categorical and continuous covariates: **Cox proportional hazards regression**
- You will see that Cox regression is a MUCH more useful method than those we have covered so far, but it makes one or two assumptions too

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)

Is PH regression a generalized linear model?

- No. As PH regression makes no assumptions about underlying distribution of survival/hazards, it is not (technically) a Generalized Linear Model (GLMs are fully parametric)
- However, it does assume that all subjects **share a common baseline hazard function** (albeit unspecified) and that any differences between survival stem **purely** from the covariates.
- A Maximum Likelihood Estimation 'related' process is still used to estimate parameters:**partial MLE**
- For this reason, partial-MLEs from PH regression can be interpreted in the same way as estimates from any GLM.
- It is only in theory (not practice) that Cox regression is not a GLM

The proportional hazards model

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

$$h_i(t) = h_0(t)e^{X\beta}$$

where
$h_i(t)$ is hazard for subject i at time t
h_0 is the baseline hazard (common to all subjects)
$X\beta$ is the subject-specific component; and
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

The proportional hazards model

$$h_i(t) = h_0(t)e^{X\beta}$$

- The covariates (in this form of the model) are time-independent: sometimes called **baseline** covariates
  - Note that there is no t component in the $X\beta$.
- Baseline covariates are typically represented by covariates such as age, sex and BMI etc.
- To consider time-varying covariates (e.g. SBP and DBP) we need to extend the PH model to include β that are dependant on t, i.e. $\beta(t)$ (we won't do this)

Hazard ratios

- The parameters estimated in Cox PH regression, β, can be transformed to obtain are hazard ratios.
- As with Logistic regression (ORs) and Poisson Regression (RRs), $\hat{\beta}$, can be **exponentiated** to get the hazard ratio (HRs). That is, $HR = e^{\beta}$
- HRs represent the risk of death (or the clinical endpoint) for one group relative to that of another (categorical covariates), or change in hazard as we increase a unit (continuous covariates).

Example: consider a single categorical covariate with two classes (Gender). If we dummy code Gender $= 0$ for males and Gender $= 1$ for females. Then...

$$h_i(t) = h_0(t)e^{\beta\, Gender}$$

Hazard ratios

If we performed a Cox PH regression and found:
$\beta_{Gender} = 2$

Then:

$$h_{males}(t) = h_0(t)e^{2\times(0)} = h_0(t)$$

$$h_{females}(t) = h_0(t)e^{2\times(1)} = h_0(t)e^2$$

$$HR = \frac{h_0(t)e^2}{h_0(t)} = e^2 = 7.4$$

This means: *Females are 7.4 time more likely to die (assuming death represents the hazard) than males*

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.

Results:

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	215	43.0%
	Censored	285	57.0%
	Total	500	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		500	100.0%

a. Dependent Variable: lenfol

Categorical Variable Codings[c]

		Frequency	(1)[b]
mitype[a]	0=non Qwave	347	1
	1=Q wave	153	0

- Of the 500 participants, 215 had heart attacks (so 285 were censored)
- I used SPSS here and (insanely) SPSS has recoded our variables to the opposite
- So Q-wave has become the referent

Results: Model 1a: (MI type)

- As we have only a single covariate, we can go straight to the coefficient table (As with any single covariate model e.g. Linear regression).

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
mitype	.660	.167	15.585	1	.000	1.935	1.394	2.685

- Significance test of β (Wald-test) gives p$<$0.05, implying difference in the hazard between the two groups
- Equivalently, the Hazard ratio, e^{β}, is significantly different from 1 (HR=1.935, p$<$0.001, 95%CI: 1.39-2.69).
- i.e. those with non-Qwave MI have 1.935 the chance of death than those with Q-wave MI.
- Equivalently, those with non-Q wave MI have a 93.5% higher

Results: Model 1b (BMI)

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
bmi	-.098	.015	44.425	1	.000	.906	.881	.933

- As in other linear models, a unit change in a continuous covariate is usually quite small (UPSHOT: wouldn't expect HRs estimates to be as profound)
- HR $= e^{\beta} = 0.906$ (p<0.001; CI:0.88-0.93)
- Interpretation: As we go up one unit of BMI, the chance of death reduces by (1-0.906)100% $= 9.4\%$.

Now a Multivariable model

- Let's try fitting both of our covariates. We will consider MI type as the study effect (effect of interest) and BMI as a potential confounder.
- We might assume a change in a coefficient (Crude$\rightarrow$Adjusted in the effect of interest of 10%) implies confounding.
- Be careful though, we should really use the change in the coefficient, not the change in the HR to gauge this (i.e. Change in coefficient $\neq$ Change in HR)

Note: In bivariate model (**Model 1a**) $\beta_{MItype} = 0.66$

Multivariable model: Model 2 (MI type and BMI)

-2 Log Likelihood	Overall (score)		
	Chi-square	df	Sig.
2393.473	56.590	2	.000

- We have >1 covariates → first test overall model (overall model is significant p<0.05)

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
bmi	-.094	.015	41.121	1	.000	.910	.884	.937
mitype	.589	.168	12.228	1	.000	1.802	1.295	2.506

- Both covariates are highly significant (p<0.05).
- Compared to $beta_{MItype} = 0.66$ from model 1, we now have $beta_{MItype} = 0.589$ (from Model 2)

$\Rightarrow 100\% \text{ x } \frac{0.66-0.589}{0.66} = 10.76\%$ change (i.e. BMI represents a confounder of the MI - survival relationship).

So what should we do?

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 Log(Hazard) linearly related to covariates
2. Hazards remain proportional over whole survival experience
   - This 2nd assumption central to Cox **Proportional Hazards** regression
   - Required to use a Maximum Likelihood like estimator to get the HR
   - Means HR 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

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

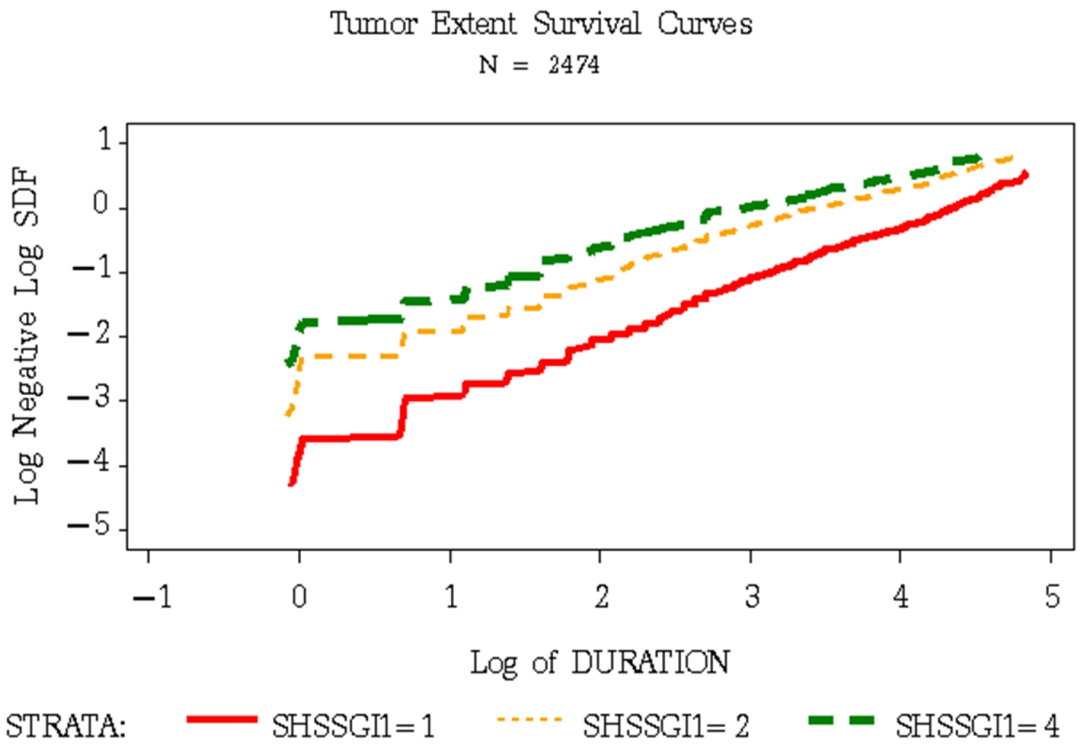
Method 1: Log minus log survival plot

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. 1997, 1999 and 2001) then the proportional hazard assumption is met.

Log minus log plots

- As the $\ln(-\ln(S(t))$ curves seem to be parallel, then the proportional hazards assumption seems to be fine



Problem with Log minus log plots:

1. Problem 1: What if we have a continuous covariate?
2. Problem 2: Subjective

Assessing the PH assumption

- The second way of assessing the PH assumption, Schoenfeld residuals, can:
  1. Consider the PH assumption for continuous covariates
  2. Provides a (formal) test of the PH assumption
- I am not a big fan of 'tests of assumptions' because they can inadvertently be under- or over-powered in studies which are otherwise meticulously designed.
- However, I consider Schoenfeld residuals an exception to the rule since considering survival curves is complicated (i.e. Not as simple as checking the histogram for normality, for example)

Disproportional hazards and difference in baseline survival

- The model we have been considering so far assumes that all individuals have the same baseline hazard
- That is, the only difference in the chance of survival between individuals (and more importantly, the underlying survival distribution) is assumed to be explained solely in terms of the covariates, AND the relative hazards are constant over the entire survival experience.
- What about if we had different groups in our data who had different underlying (baseline) distributions of survival?

Disproportional hazards and difference in baseline survival

When might this occur?

1. If we are considering two populations that might differ in many ways (specifically in ways that are difficult to measure or articulate in a model)
   - For example, if we are considering both the Australian and Thai populations, there are differences in cultural practice, lifestyle, diet and genetics.
2. If our proportional hazard assumption isn't met (for a particular covariate).
   - For example, what about if the hazard ratio (of males relative to females) was 2 at day 50 (twice the chance of death), but this increased to 5 at 200 days. The hazard ratio is clearly not constant over the entire survival experience.

Dealing with different baseline survival $\rightarrow$ hazards

If this occurs for a single covariate only (preferably categorical), we can used Stratified Cox regression model
In this model, a different baseline hazard is allowed for each of stratum, taking the form:

$$h(t|s_i) = h0_i(t)e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

for stratum $i = 1, 2, \cdots, m$ strata

- **Note:** While the baseline hazards is allowed to differ for each stratum; the effect of the covariates (on survival) remains the same (i.e the βs do not vary across strata)
- For that reason, the output from a stratified Cox regression is no more complicated than a standard Cox regression

Example of a stratified Cox regression

- The output looks exactly the same. The only difference is that that 'populations' are allowed to have different baseline survival curves.
- In this case, I have stratified by (cohort) year (1997, 1999, 2001)

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
bmi	-.093	.015	39.122	1	.000	.911	.885	.938

Your turn: How would you interpret the hazard ratio?