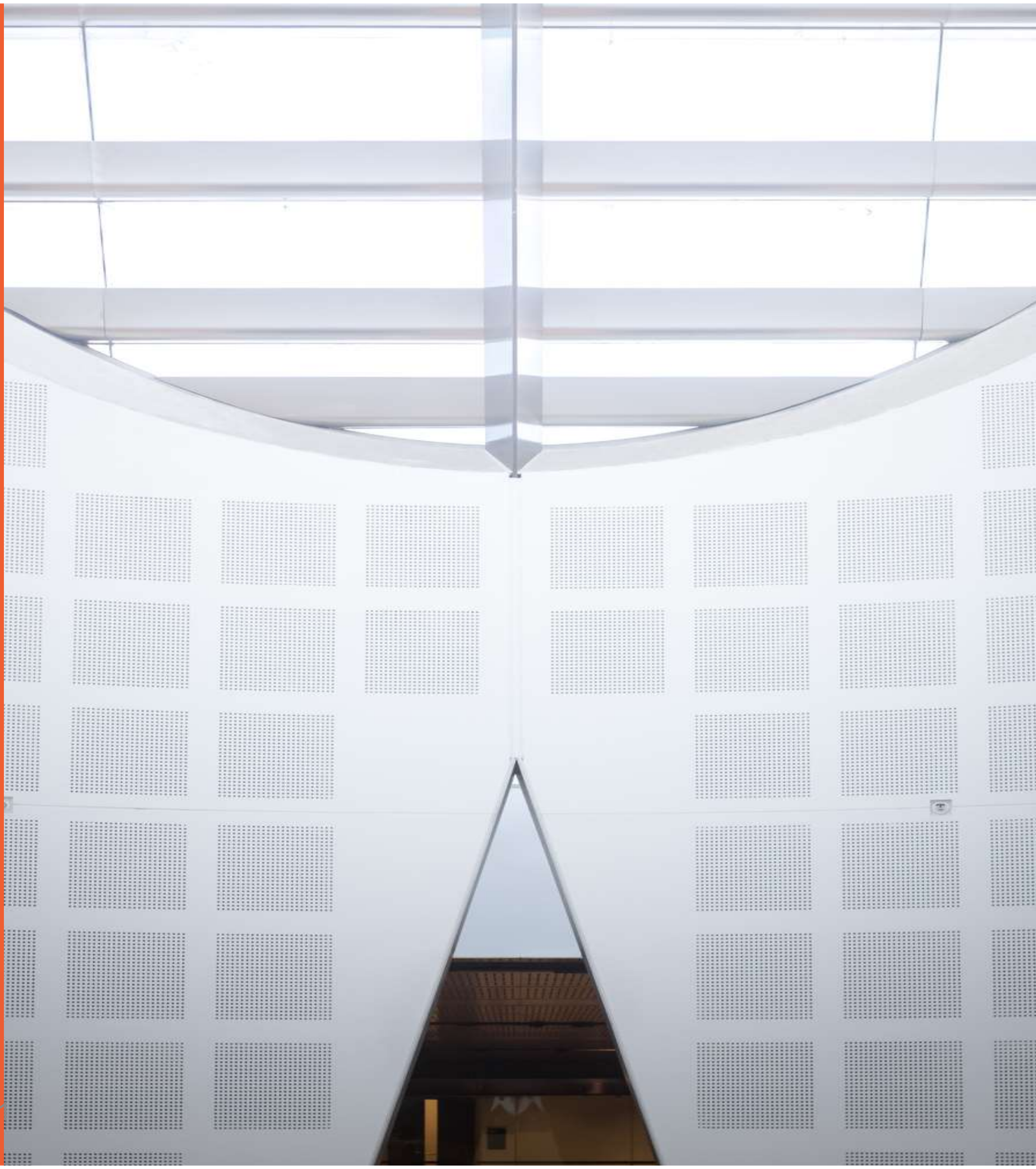
Introduction to Survival Analysis

Presented by
Dr Kathrin Schemann
Sydney Informatics Hub
Core Research Facilities
The University of Sydney

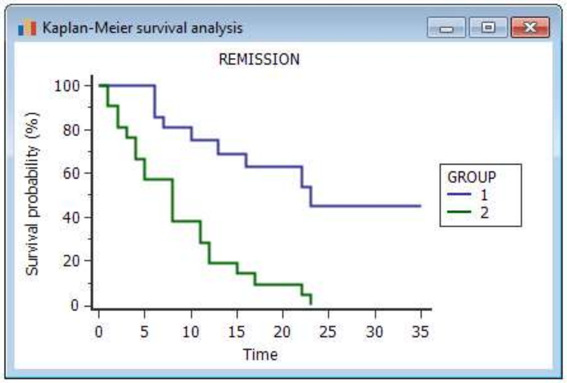


Outline

- Survival Analysis – brief description
- Kaplan Meier – description: how it works
- Kaplan Meier – workflow: how to do the analysis (with worked example)
- Cox proportional hazards regression – description
- Cox PH regression – workflow: how to do the analysis (with worked example)
- Other varieties of survival analysis
- Software options and references

Workshop Aims

- Understand the key concepts in Survival Analysis



- Follow the steps to perform Kaplan-Meier and Cox Regression

How to use this workshop

- These slides have a dual purpose:
  - To guide our interactive workshops
  - As self-contained reference material and workflows to be used after the workshop

- Some slides are for your reference, and not all of the material will be discussed in the workshop. Such slides are marked with this blackboard icon

- Ask short questions or clarifications during the workshop. There will be breaks during the workshop for longer questions. You can email us about the material in these workshops at any time, or request a consultation for more in-depth discussion of the material as it relates to your specific project.

How to use this workshop

Reference Information

The primary example used in this workshop comes from the book "Applied Survival Analysis" 2nd Ed, by Hosmer and Lemeshow.

https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/1367smt/scopus2-s2.0-84947789021

Download the data as a zip file from: ftp://ftp.wiley.com/public/sci_tech_med/survival

The files to use are whas500.dat and whas500.txt

The SPSS syntax used for workshop examples is also available to workshop participants.

An R Markdown file and R script covers how to do the equivalent analyses in R.

Introduction

When to use Survival Analysis?

- When you measure the time elapsed until a specified event occurs.
- The event doesn't have to occur for all subjects. This is an important feature of survival analysis
- The classic event is "death" which gives survival analysis its name.

Alternative analysis:

- Logistic regression models the probability of the event occurring within a timeframe, not the rate over time.

Survival Time and Event

Examples

Description of survival time	Event
Overall Survival – time a person lives after cancer surgery	death
Progression Free Survival - time to progression or death from any cause	death/progression
Remission – time a person is disease free since cancer treatment	relapse
Machine Reliability - Duration that a machine operates without fault	failure
Fertility – Duration from fertility treatment to pregnancy and subsequent birth	birth
Churn – time a household spends with an internet service provider	switch provider

↑

Event must be definitive

Survival Analysis models and tests

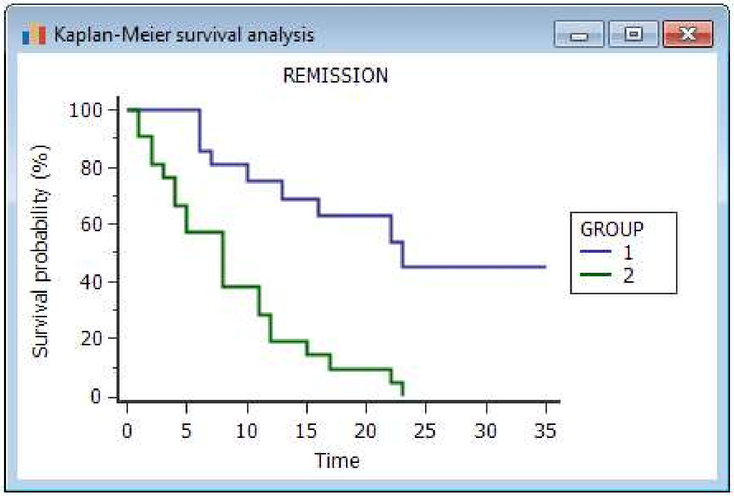
1. Kaplan-Meier – “non-parametric” meaning that there is no assumption about the shape of the survival curve.
2. Cox proportional hazards regression – this is the most common model that we think of in survival analysis (it is semi-parametric)
3. Parametric regression models – like Cox, but assumes an underlying survival distribution (eg Exponential, Weibull, etc)
4. Frailty models – allows clustering to be modelled with a random effect (like in Mixed Models)
5. Competing Risks models – partitions event types
6. Discrete Time model using logistic regression – used when time is measured discretely with only a few values possible

Kaplan-Meier Introduction

The Kaplan-Meier procedure is commonly used to estimate the survival function, *S*(*t*).

S(*t*) represents the probability of observing a survival time greater than time, *t*.

We use the observed data to estimate the conditional probability of confirmed survival at each observed survival time and then multiply them to obtain an estimate.



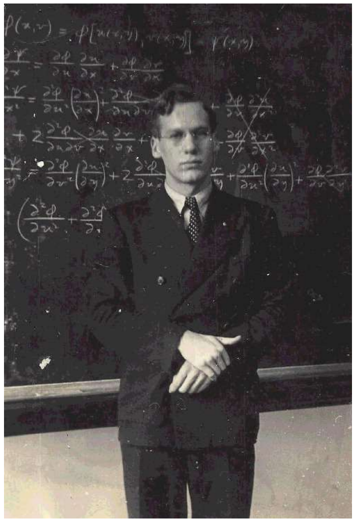
Kaplan-Meier Introduction

Did you know?

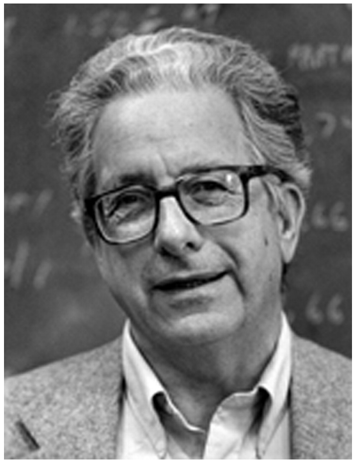
Edward Kaplan and Paul Meier worked on survival separately in the 1950's and submitted separate papers to JASA in ~1954. Their mentor, John Tukey, got them together and the work was jointly published in 1958.

Their estimator for the survival curve became known as the Kaplan-Meier method which became the standard way to report patient survival data in medical research.

Their paper is the eleventh most cited scientific paper of the modern era (@ 2014).



Kaplan
-
Meier



Kaplan-Meier Introduction

Kaplan-Meier estimator of the survival function

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

n_i = number at risk of dying at time of *ith* observed event

d_i = number of observed deaths at the *ith* observed event

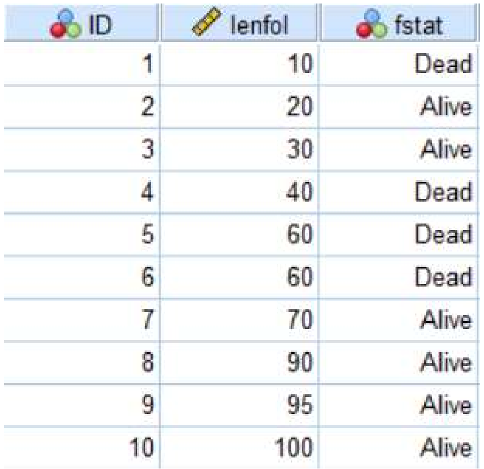
$\frac{n_i - d_i}{n_i}$ = probability of surviving at the *ith* observed event

$\hat{S}(t)$ =1 at the time origin, t=0

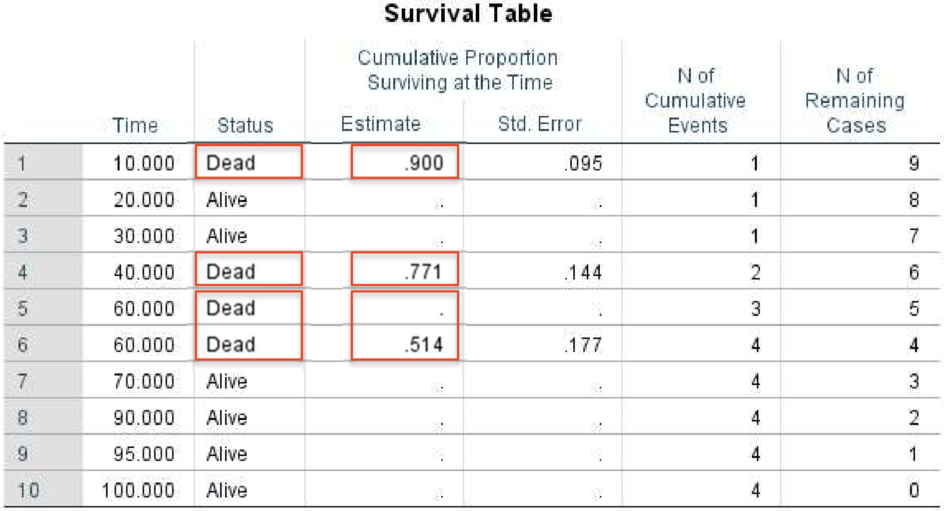
At any point in time *S(t)* is estimated by multiplying a sequence of conditional survival probability estimators.

Kaplan-Meier Introduction

Illustration of Survival function: example (SPSS)



ID	lenfol	fstat
1	10	Dead
2	20	Alive
3	30	Alive
4	40	Dead
5	60	Dead
6	60	Dead
7	70	Alive
8	90	Alive
9	95	Alive
10	100	Alive



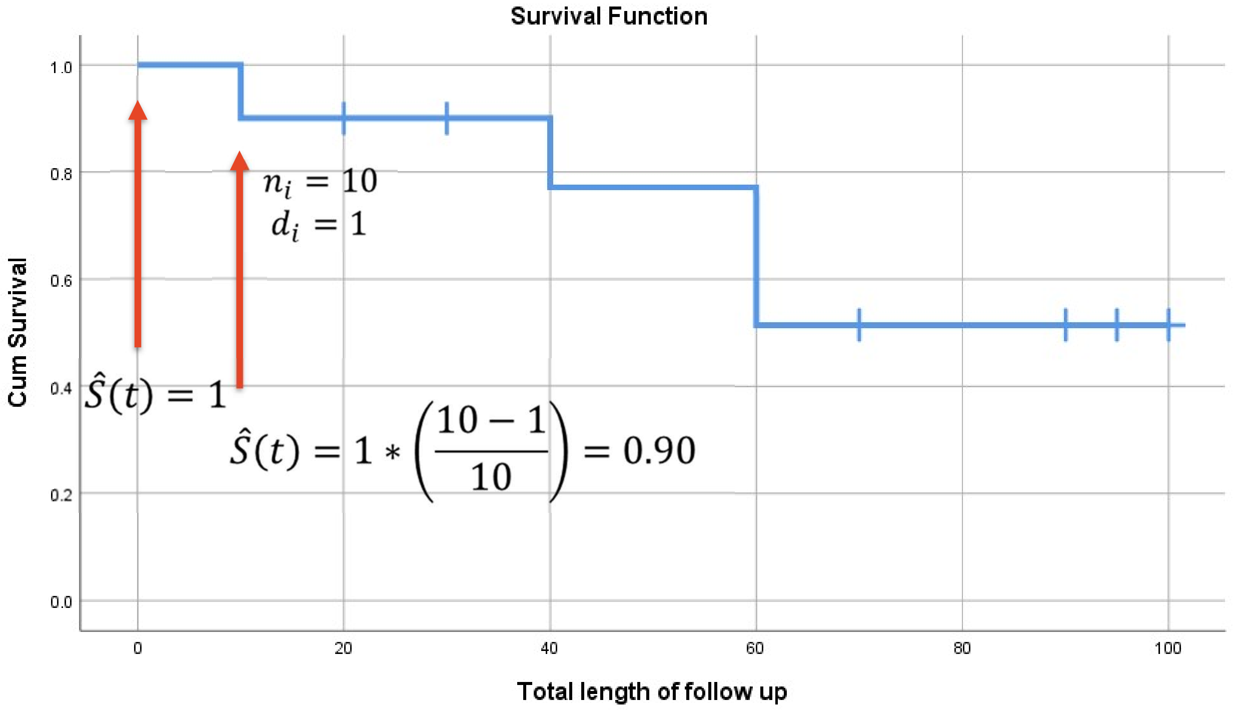
Survival Table

			Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
	Time	Status	Estimate	Std. Error		
1	10.000	Dead	.900	.095	1	9
2	20.000	Alive	.	.	1	8
3	30.000	Alive	.	.	1	7
4	40.000	Dead	.771	.144	2	6
5	60.000	Dead	.	.	3	5
6	60.000	Dead	.514	.177	4	4
7	70.000	Alive	.	.	4	3
8	90.000	Alive	.	.	4	2
9	95.000	Alive	.	.	4	1
10	100.000	Alive	.	.	4	0

The survival table shows the value of the survival function changing at each timepoint when an event or events occur.

Kaplan-Meier Introduction

Illustration of Survival function: example

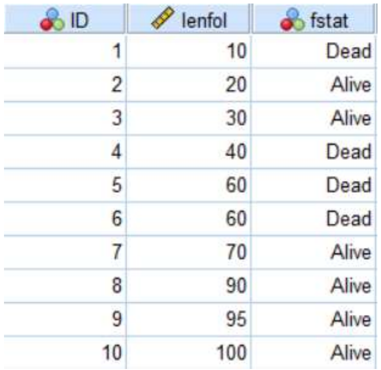


$n_i = 10$

$d_i = 1$

$\hat{S}(t) = 1$

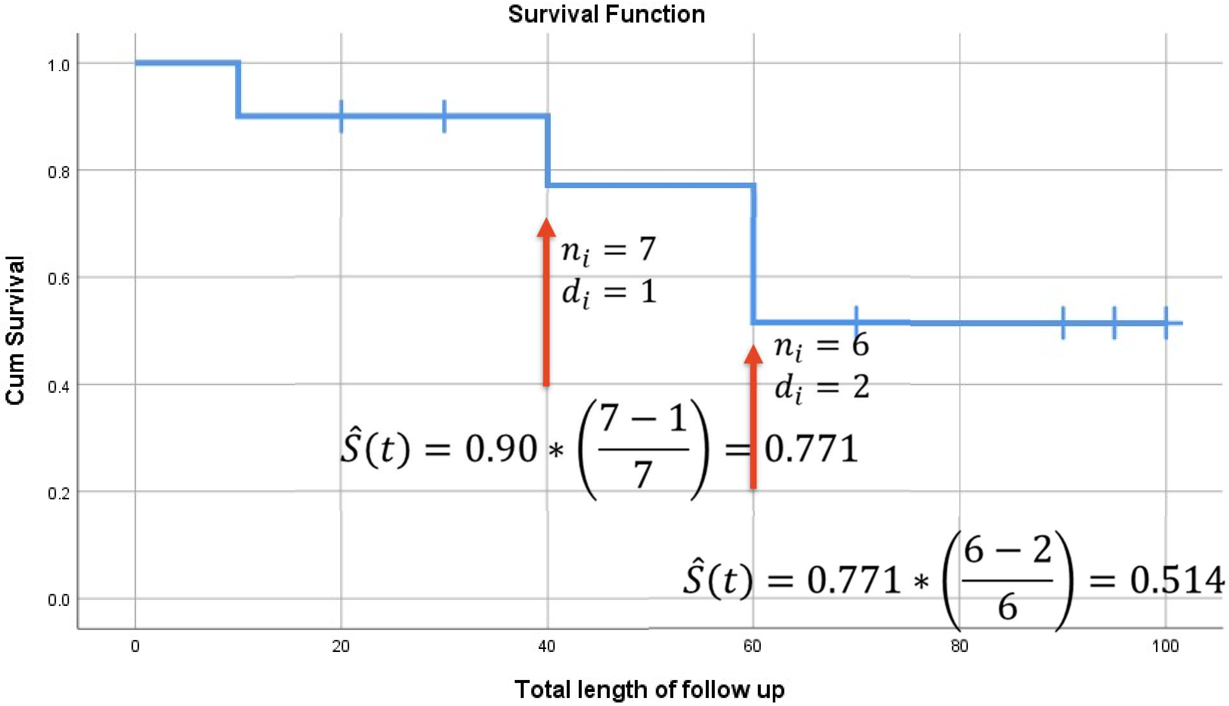
$$\hat{S}(t) = 1 * \left(\frac{10 - 1}{10}\right) = 0.90$$



ID	lenfol	fstat
1	10	Dead
2	20	Alive
3	30	Alive
4	40	Dead
5	60	Dead
6	60	Dead
7	70	Alive
8	90	Alive
9	95	Alive
10	100	Alive

Kaplan-Meier Introduction

Illustration of Survival function: example

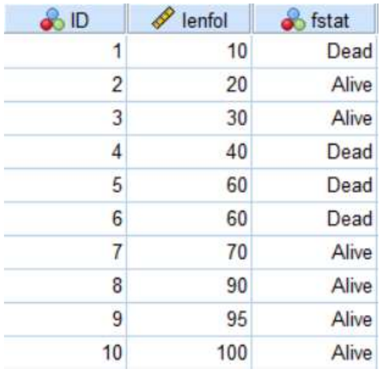


$n_i = 7$
$d_i = 1$

$n_i = 6$
$d_i = 2$

$$\hat{S}(t) = 0.90 * \left(\frac{7-1}{7}\right) = 0.771$$

$$\hat{S}(t) = 0.771 * \left(\frac{6-2}{6}\right) = 0.514$$



ID	lenfol	fstat
1	10	Dead
2	20	Alive
3	30	Alive
4	40	Dead
5	60	Dead
6	60	Dead
7	70	Alive
8	90	Alive
9	95	Alive
10	100	Alive

Kaplan-Meier workflow

With the Kaplan-Meier procedure we can plot the survival curves for an event and compare a single factor

1. Data - Define the time variable, the event variable and any nominal explanatory variables
2. Procedure - Run the K-M procedure in your software to produce survival descriptive statistics and plots, and test statistics such as Log-rank.
3. Interpretation - Interpret the results

Kaplan-Meier 1. Data

Time to Event

- What is the event? Make sure it is binary.
- How do we define the time to it?
- Define the beginning and end points.

Types of observations

1. The event occurred and we measure when it occurred
2. The event did not occur within a known time period

Explanatory variables

– Record nominal variables of interest

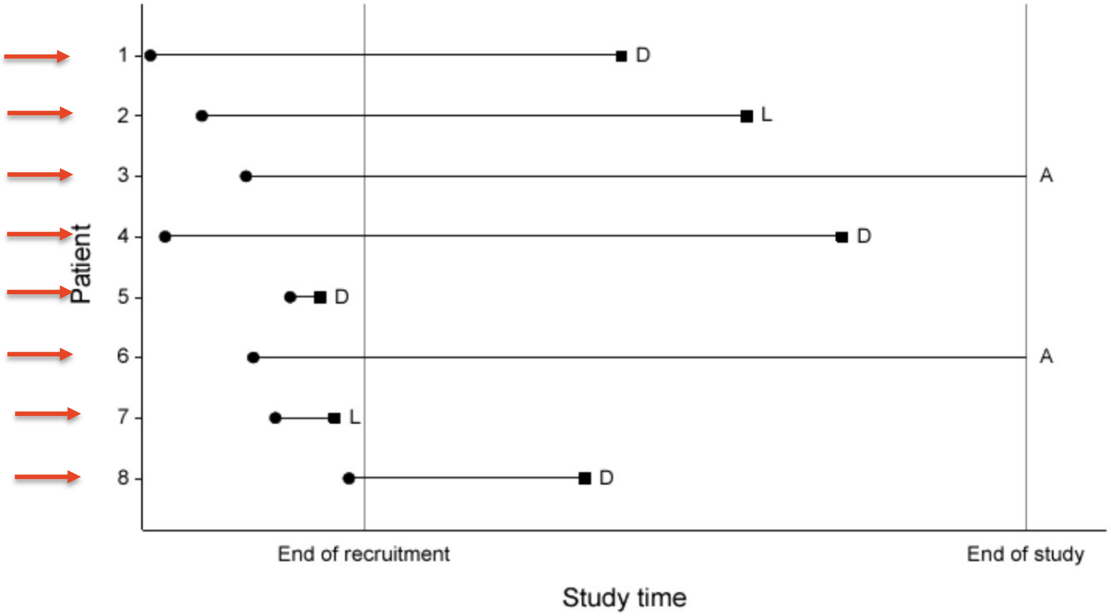
Kaplan-Meier 1. Data
Censoring

Censoring just means that we are missing some information of interest. It can have different causes.

1. A subject has not experienced the event during the study period
2. A subject is lost to follow up during the study period
3. A subject experiences a different event that makes further follow up impossible.

Kaplan-Meier 1. Data

Censoring



Patients 1, 4, 5 & 8 die and their survival time is recorded

Patients 2 & 7 are lost to follow up – right censored

Patients 3 & 6 are alive at the end of the study – right censored

Kaplan-Meier 1. Data Assumptions

Assumption 1: censoring is independent (non-informative)

This means for example that loss to follow up is not associated with a higher probability of the event occurring.

Assumption 2: Survival probability is independent on when a subject enters the study (recruitment often occurs over a period of time).

Assumption 3: The event occurs at the time it is recorded.

This is relevant when the observation of the event occurs during a follow up visit for example.

Kaplan-Meier 1. Data

Example: Worcester Heart Attack Study

The goal of the study was to study factors and time trends associated with long-term survival following acute myocardial infarction (MI) among residents of Worcester, Massachusetts, USA.

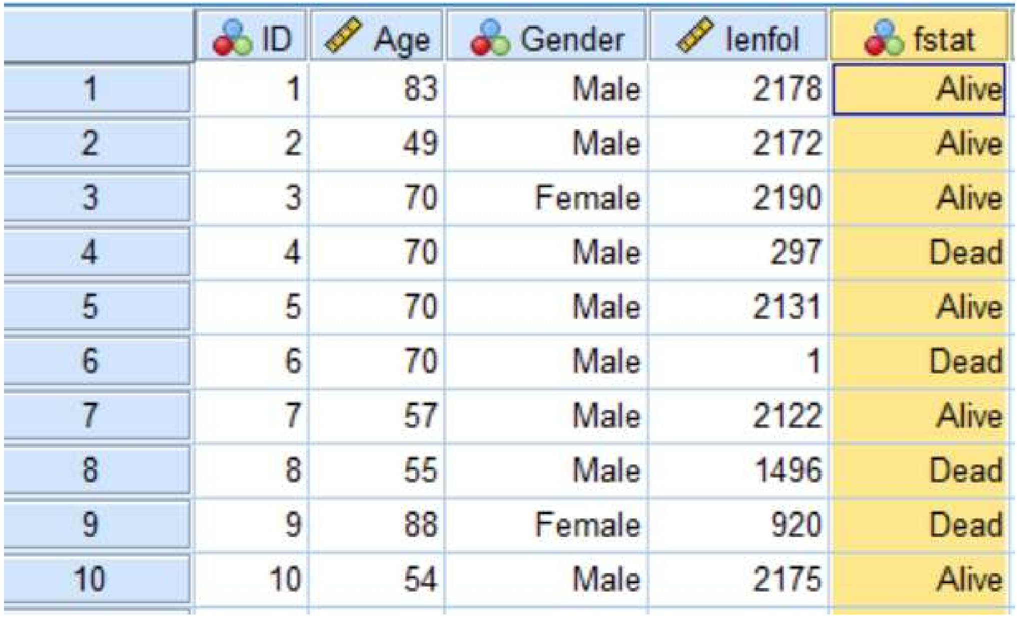
(reference: Applied Survival Analysis 2nd Ed)

What is the event?	Death (due to any cause)
Time to event?	From hospital admission date to date of last follow up (in days)

We will consider Sex (Gender) as a factor.

Kaplan-Meier 2. Procedure

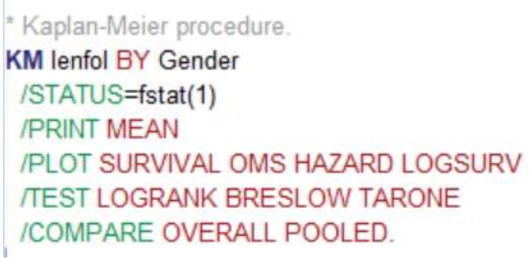
First 10 rows of data



	ID	Age	Gender	lenfol	fstat
1	1	83	Male	2178	Alive
2	2	49	Male	2172	Alive
3	3	70	Female	2190	Alive
4	4	70	Male	297	Dead
5	5	70	Male	2131	Alive
6	6	70	Male	1	Dead
7	7	57	Male	2122	Alive
8	8	55	Male	1496	Dead
9	9	88	Female	920	Dead
10	10	54	Male	2175	Alive

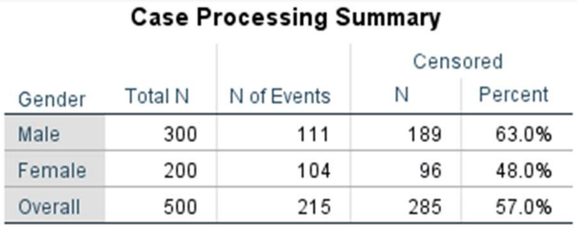
Kaplan-Meier 2. Procedure

Run procedure in your chosen software e.g. SPSS



```
* Kaplan-Meier procedure.
KM lenfol BY Gender
  /STATUS=fstat(1)
  /PRINT MEAN
  /PLOT SURVIVAL OMS HAZARD LOGSURV
  /TEST LOGRANK BRESLOW TARONE
  /COMPARE OVERALL POOLED.
```

Have a look at the number of events in the dataset and the percentage censored.

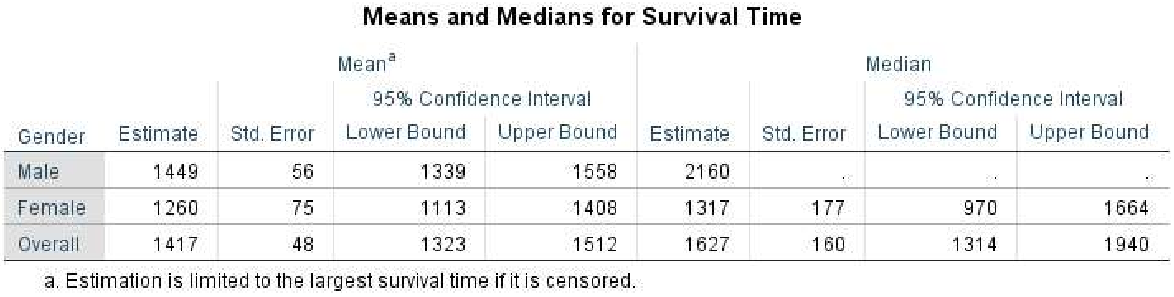


Case Processing Summary

			Censored	
Gender	Total N	N of Events	N	Percent
Male	300	111	189	63.0%
Female	200	104	96	48.0%
Overall	500	215	285	57.0%

Q: Why do we want to look at the Case Processing Summary?

Kaplan-Meier 2. Procedure



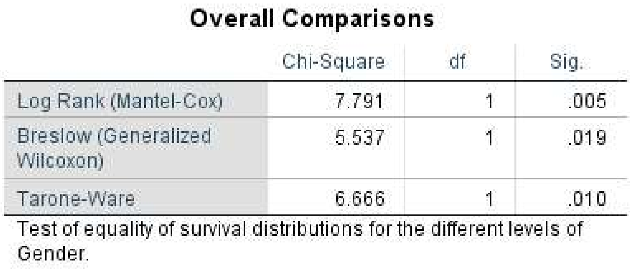
Means and Medians for Survival Time

	Mean[a]				Median			
			95% Confidence Interval				95% Confidence Interval	
Gender	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
Male	1449	56	1339	1558	2160	.	.	.
Female	1260	75	1113	1408	1317	177	970	1664
Overall	1417	48	1323	1512	1627	160	1314	1940

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

Test for difference between Male and Female

Log-rank, Breslow and Tarone-Ware statistics are all significant



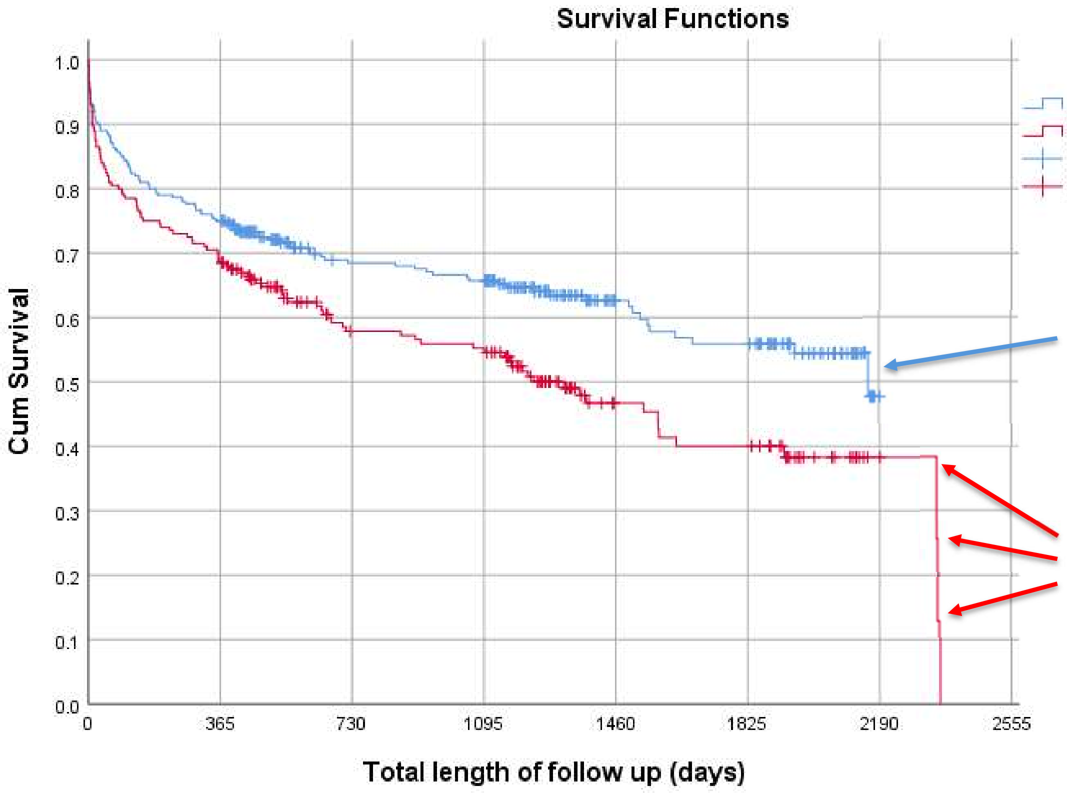
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	7.791	1	.005
Breslow (Generalized Wilcoxon)	5.537	1	.019
Tarone-Ware	6.666	1	.010

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

The log-rank test calculates the difference between the observed events for each group with the expected events for the combined groups and weights timepoints equally. The Breslow test weights timepoints according to number at risk, n_i, while Tarone-Ware weights timepoints by $\sqrt{n_i}$.

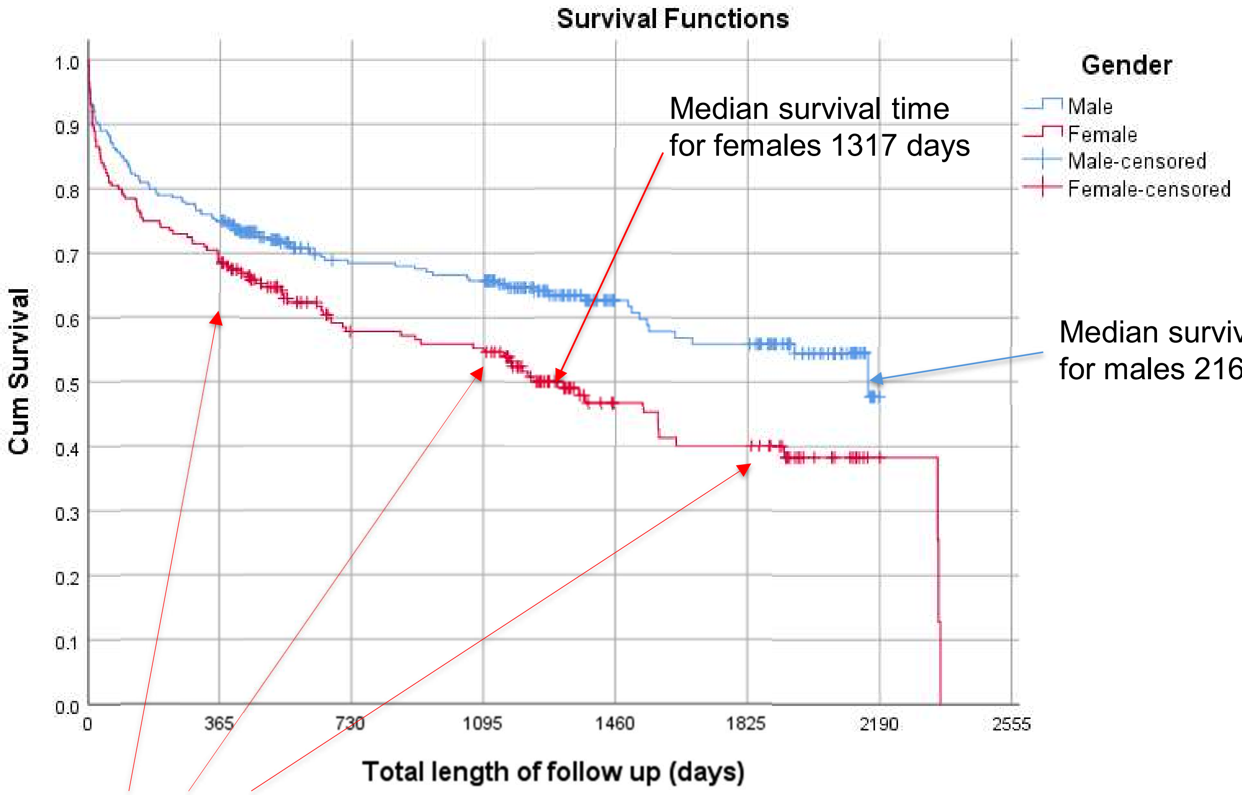
Kaplan-Meier 2. Procedure



Final event for males at 2160 days, with n_i = 8

Final 3 events for females at 2350, 2353 & 2358 days, with n_i = 3,2,1

Kaplan-Meier 2. Procedure



Median survival time for females 1317 days

Median survival time for males 2160 days

1yr, 3yr & 5yr follow up times

Kaplan-Meier 3. Interpretation

There is a significant difference in survival between males and females (by log-rank test)

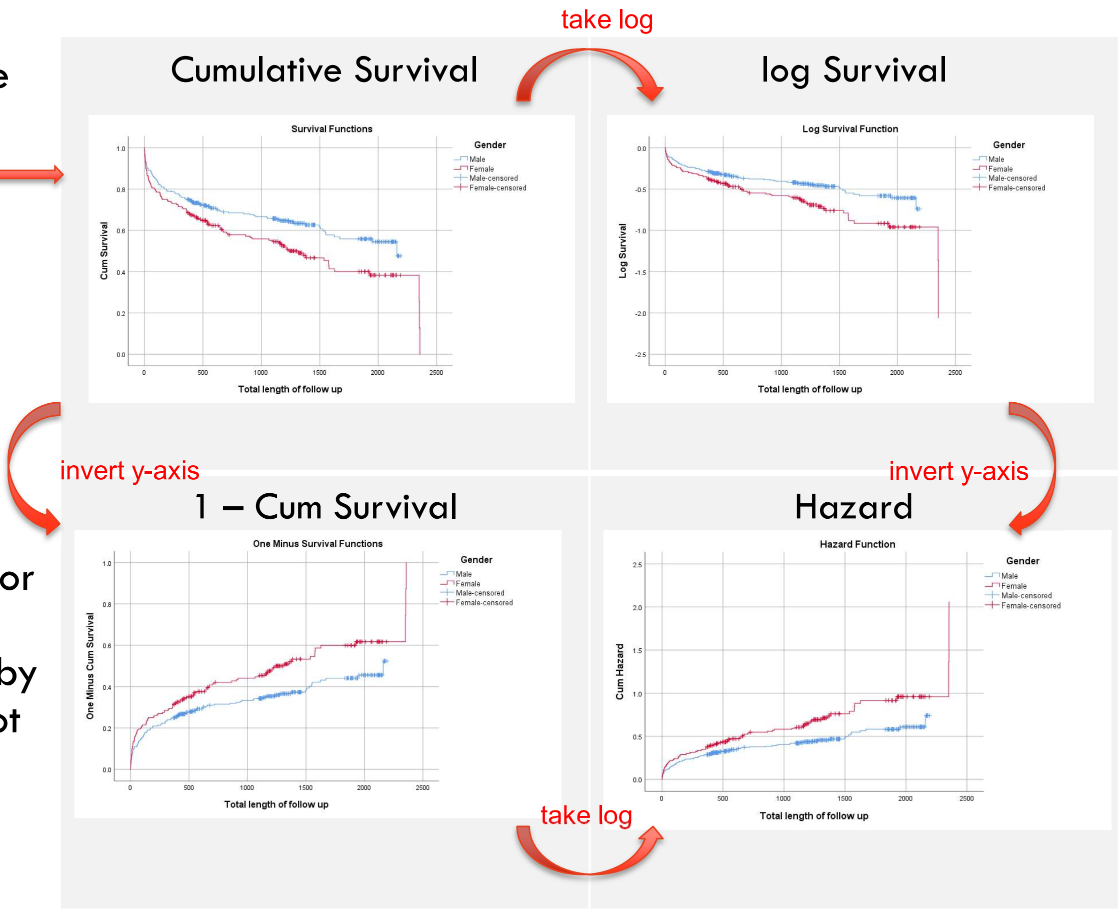
Median survival for males: 2160 days [95%CI: not calc]

Median survival for females: 1317 days [95% CI 970-1664]

Why didn’t we get a CI for males?
Because the last event occurred before we hit 50% survival.

Kaplan-Meier 3. Interpretation

Include the cumulative Survival curve plot in your report.



take log

Cumulative Survival

log Survival

invert y-axis

invert y-axis

1 – Cum Survival

Hazard

take log

The Survival function or Hazard function may also be represented by a variety of other plot options (as shown).

Kaplan-Meier
Challenge questions

Q1: Which of the following statements is not correct:

The Kaplan-Meier method can be used when

a) There is a binary outcome dead/alive
b) Subjects are observed over time
c) You want to estimate a hazard rate for survival
d) There is no more than one categorical factor of interest

Kaplan-Meier
Challenge questions

Q1: Which of the following statements is not correct:

The Kaplan-Meier method can be used when

a) There is a binary outcome dead/alive

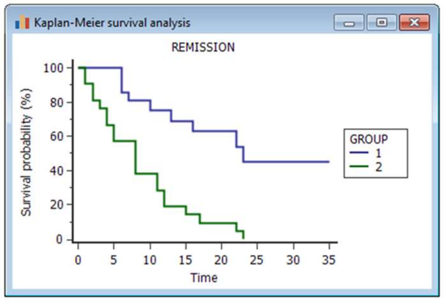
b) Subjects are observed over time

[✓ c) You want to estimate a hazard rate for survival]

d) There is no more than one categorical factor of interest

Kaplan-Meier

Any questions?



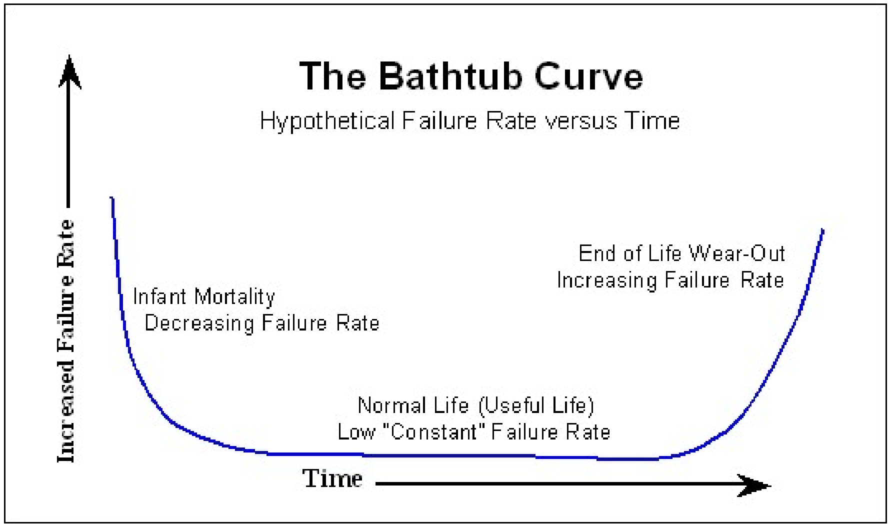
Cox Proportional Hazards Regression Introduction

- Semi-parametric model: Does not assume an underlying distribution of survival time. Dependence on time is unspecified
- Covariates are parameterised in a similar way to linear regression. Their value must remain constant over time.
- The baseline hazard function is like the intercept in linear regression
- The covariate parameter estimates are called Hazard Ratios and are similar to Odds Ratios in logistic regression
- The proportional hazards assumption allows us to interpret the HR's as a constant over time. This needs to be checked.

Cox Proportional Hazards Regression Introduction

Semi-parametric model: Does not assume an underlying distribution of survival time. Dependence on time is unspecified.

Note: Cox Hazard function is constant over time. This is not always true. For human life the function is bathtub shaped. High in perinatal period, then low for a long time, then high again.



Cox Proportional Hazards Regression Introduction

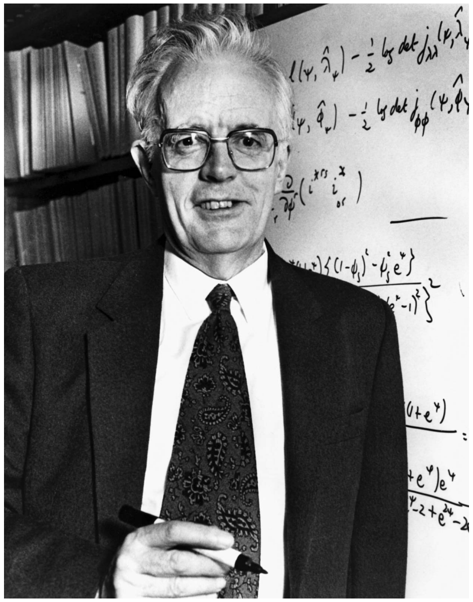
Did you know?

The Cox Regression method was developed by David Cox (British statistician) based on the earlier Kaplan-Meier work.

He cited the KM paper in 1959. The second of only 11 citations for KM over the first 11years following publication!

Both Kaplan-Meier and Cox regression took off after the publication of his paper in 1972.

He died on 18th January 2022, aged 97.



Cox Proportional Hazards Regression Workflow

1. What do we want to model?
2. EDA – Look at Kaplan-Meier survival curve
3. Build the Cox regression model
4. Check the model assumptions
5. Interpret the model

Cox Proportional Hazards Regression Workflow

1. What do we want to model?

Example: Worcester Heart Attack Study (WHAS)

As before, the event is death (due to any cause).

There may be many potential explanatory factors that we wish to examine, for example:

- Gender
- Age (at admission)
- Initial heart rate
- Initial systolic blood pressure
- Initial diastolic blood pressure
- BMI
- History of cardiovascular disease
- Atrial fibrillation
- Cardiogenic Shock
- Congestive heart complications
- Complete heart block
- MI order
- MI type
- Cohort Year

Cox Proportional Hazards Regression Workflow

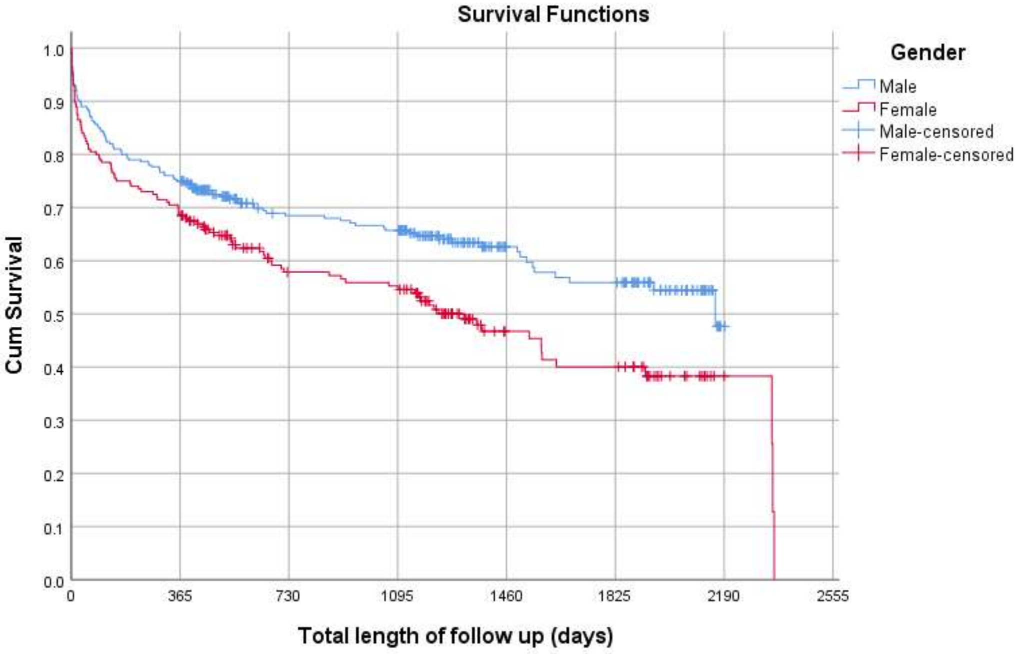
1. What do we want to model?

Let’s start with a simple univariate model including Gender (Sex), then we will build more complex models.

Cox Proportional Hazards Regression Workflow

2. EDA – Kaplan-Meier curve

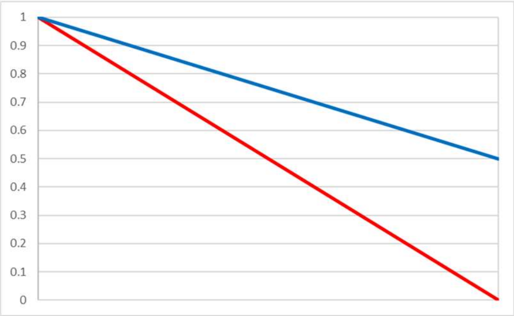
Have a look at the primary covariate of interest. Do the curves look proportional over the study period?



Cox Proportional Hazards Regression Workflow

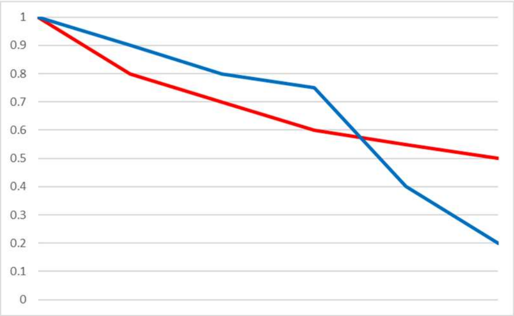
2. EDA – Kaplan-Meier survival curve

Proportional hazards



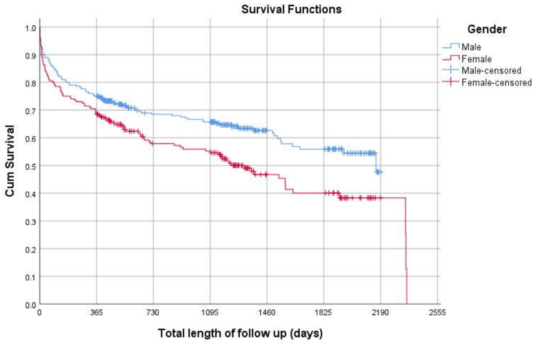
Constant rates for m & f
Unchanging Hazard Ratio
over time

Meets assumption,
but not real life!



Non-proportional
Hazard Ratio changes over
time from <1 to >1 for m:f

Fails assumption of
proportional hazards



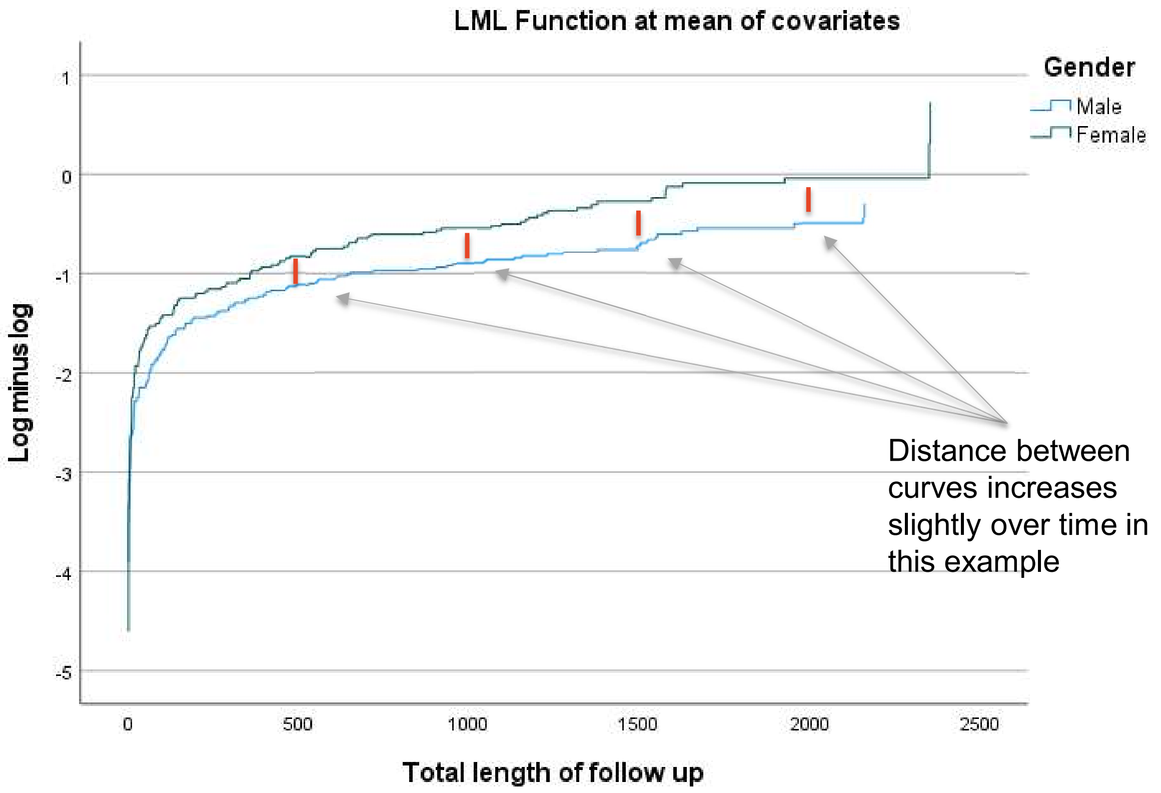
Changing rates but
Hazard Ratio appears
stable over time

Meets assumption of
proportional hazards

Cox Proportional Hazards Regression Workflow

2. EDA – Log minus Log curve

Another common method of checking the PH assumption is to plot a transformation of the survival data known as “log minus log”. If the PH assumption is met, then the two curves will be equidistant apart along the length.



Distance between curves increases slightly over time in this example

Note: In SPSS this plot is created in the Cox procedure using the factor as a “strata”variable. It is not available in the K-M procedure.

Cox Proportional Hazards Regression Workflow

3. Model Fitting

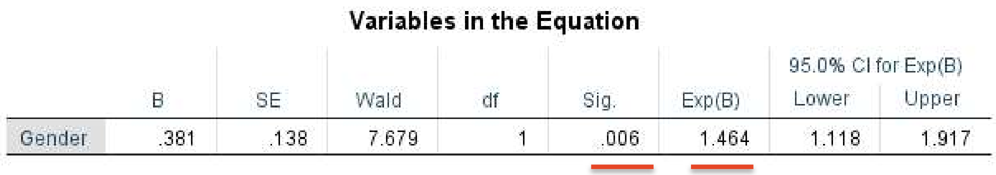
Basic techniques are identical to those used in logistic regression

- Maximum Likelihood methods used to obtain parameter estimates and SE's
- Use (partial) log-likelihood and chi square test to assess overall significance and compare nested models.
- Check for significance of interaction terms

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Run the Cox Regression procedure using your chosen software (SPSS shown)



Variables in the Equation

							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Gender	.381	.138	7.679	1	.006	1.464	1.118	1.917

Gender is significant. We keep it in the model.

Hazard Ratio (Gender) = 1.464

(The odds of death occurring first for a female is ~1.5 compared to a male)

Cox Proportional Hazards Regression Workflow

3. Model Fitting

What about all the other covariates of interest?

The choice of strategy for model building is similar to those used in ordinary linear regression

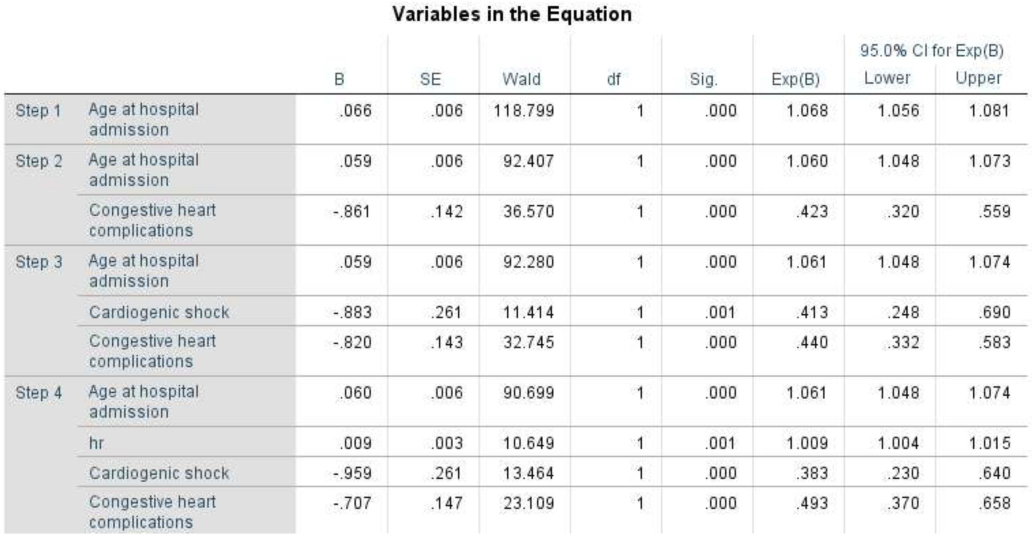
	Hierarchical	Simultaneous	Stepwise
Style	most academic		least academic
Theory	Strong theory	Limited theory	no theory
Analyst role in model building	choose variables, and the order of entry	choose a list of variables believed to be important	Variables are chosen through automated process
Possible use	Designed experiments	Exploratory	Data mining type approach

Cox Proportional Hazards Regression Workflow

3. Model Fitting

We have 14 variables to explore for their association with survival.

Simultaneous approach – using a method that applies a (likelihood ratio) test to decide which variables to keep in the model.
Here are the first 4 steps in SPSS…



Variables in the Equation

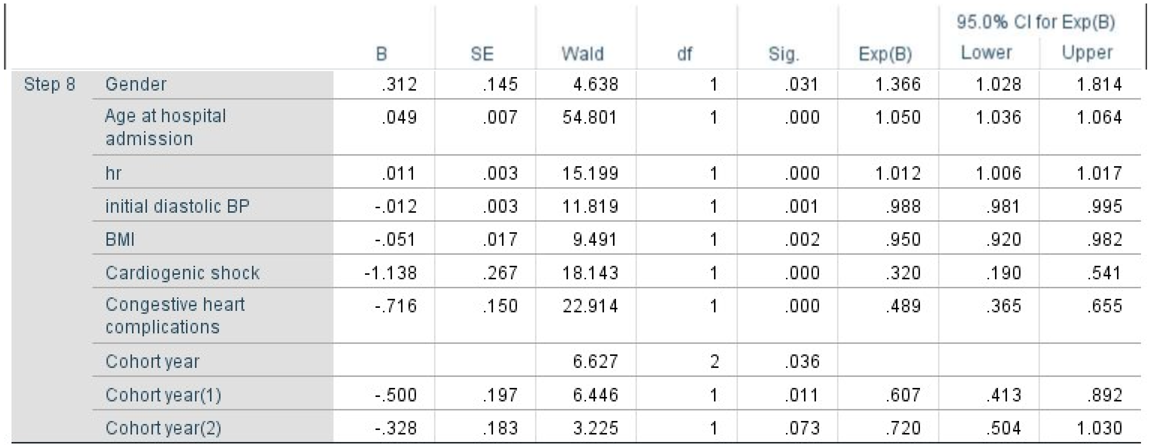
		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	Age at hospital admission	.066	.006	118.799	1	.000	1.068	1.056	1.081
Step 2	Age at hospital admission	.059	.006	92.407	1	.000	1.060	1.048	1.073
	Congestive heart complications	-.861	.142	36.570	1	.000	.423	.320	.559
Step 3	Age at hospital admission	.059	.006	92.280	1	.000	1.061	1.048	1.074
	Cardiogenic shock	-.883	.261	11.414	1	.001	.413	.248	.690
	Congestive heart complications	-.820	.143	32.745	1	.000	.440	.332	.583
Step 4	Age at hospital admission	.060	.006	90.699	1	.000	1.061	1.048	1.074
	hr	.009	.003	10.649	1	.001	1.009	1.004	1.015
	Cardiogenic shock	-.959	.261	13.464	1	.000	.383	.230	.640
	Congestive heart complications	-.707	.147	23.109	1	.000	.493	.370	.658

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Stepwise:

Here is the final step



		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 8	Gender	.312	.145	4.638	1	.031	1.366	1.028	1.814
	Age at hospital admission	.049	.007	54.801	1	.000	1.050	1.036	1.064
	hr	.011	.003	15.199	1	.000	1.012	1.006	1.017
	initial diastolic BP	-.012	.003	11.819	1	.001	.988	.981	.995
	BMI	-.051	.017	9.491	1	.002	.950	.920	.982
	Cardiogenic shock	-1.138	.267	18.143	1	.000	.320	.190	.541
	Congestive heart complications	-.716	.150	22.914	1	.000	.489	.365	.655
	Cohort year			6.627	2	.036			
	Cohort year(1)	-.500	.197	6.446	1	.011	.607	.413	.892
	Cohort year(2)	-.328	.183	3.225	1	.073	.720	.504	1.030

This is one approach to quickly explore and discover possible explanatory factors that meet the significance threshold.

There are many caveats to this process! Please attend **"Statistical Model Building"** to gain a more complete understanding of issues involved.

Cox Proportional Hazards Regression Workflow

3. Model Fitting - Collett

Suppose we have a list of variables believed to be important, or that we need to control for.

- Age
- Gender
- BMI
- Heart Rate

We can use a more rigorous model building approach described by David Collett in “Modelling Survival Data in Medical Research”

Cox Proportional Hazards Regression Workflow

3. Model Fitting - Collett

Collett's general strategy for model selection:

1. Fit univariate models for each predictor variable of interest. Compare the -2loglikelihood values to the null model (using chi square statistic, or AIC, BIC)
2. Significant variables from step 1 are fitted in a single model and compared to 'leave one out' models
3. Variables omitted at step 1 are then added to the best model from step 2 and compared
4. A final check to ensure no term in the model can be omitted without increasing -2LL significantly and no new term added without reducing -2LL significantly.

Cox Proportional Hazards Regression Workflow

3. Model Fitting - Collett

Collett's steps for model selection:

model no.	Variables in the model	-2 log L	AIC -2LogL+2q
1	Null	2455.2	2455.2
2	Age	2313.4	2315.4
3	Gender	2447.6	2449.6
4	BMI	2407.0	2409.0
5	Heart Rate	2426.3	2428.3
6	Age + Gender +BMI	2305.5	2311.5
7	Age + Gender + HR	2294.5	2300.5
8	Age + BMI + HR	2287.9	2293.9
9	Gender + BMI + HR	2379.4	2385.4
10	Age+Gender+BMI+HR	2286.8	2294.8

1. Chi-sq significant, all variables go to next step

2. Compared to the full model 10, model 8 has a lower AIC.

Cox Proportional Hazards Regression Workflow

3. Model Fitting - Collett

Collett's steps for model selection:

model no.	Variables in the model	-2 log L	AIC -2LogL+2q
1	Null	2455.2	2455.2
2	Age	2313.4	2315.4
3	Gender	2447.6	2449.6
4	BMI	2407.0	2409.0
5	Heart Rate	2426.3	2428.3
6	Age + Gender +BMI	2305.5	2311.5
7	Age + Gender + HR	2294.5	2300.5
8	Age + BMI + HR	2287.9	2293.9
9	Gender + BMI + HR	2379.4	2385.4
10	Age+Gender+BMI+HR	2286.8	2294.8

No variables from step 1 were omitted. Steps 3 & 4 not required.

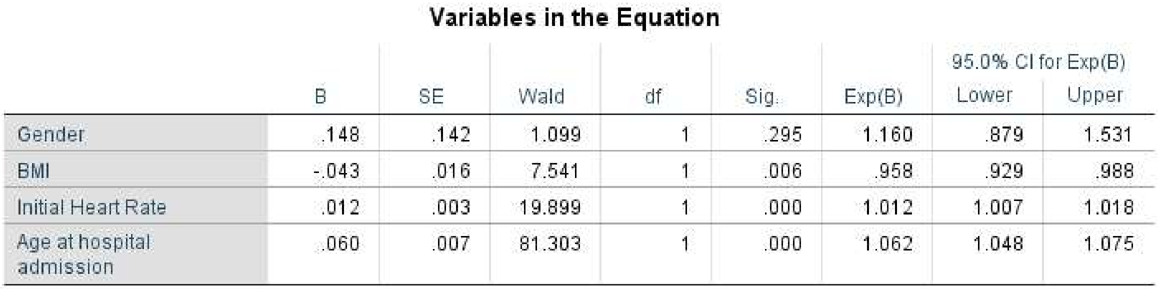
Lowest AIC: model 8

Cox Proportional Hazards Regression Workflow

3. Model Fitting - Collett

Model 8 gave lowest AIC, but I want to report on Gender, so I will include that in the model as well.

Model 10: Parameter Estimates



Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Gender	.148	.142	1.099	1	.295	1.160	.879	1.531
BMI	-.043	.016	7.541	1	.006	.958	.929	.988
Initial Heart Rate	.012	.003	19.899	1	.000	1.012	1.007	1.018
Age at hospital admission	.060	.007	81.303	1	.000	1.062	1.048	1.075

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Other points to check during covariate selection

- Linearity of continuous predictors (eg BMI)
- Interactions between predictors
- Avoid Overfitting (have at least 10 events per covariate df)

Worcester Heart Attack Study

- We have 215 events. Plenty of statistical power to include many predictors

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Linearity of continuous predictors (eg BMI)

Based on US and Aust Gov’t health guidelines, we classify BMI into ranges:

<18.5	Underweight
18.5-24.9	Normal weight
25-29.9	Overweight
≥30	Obese

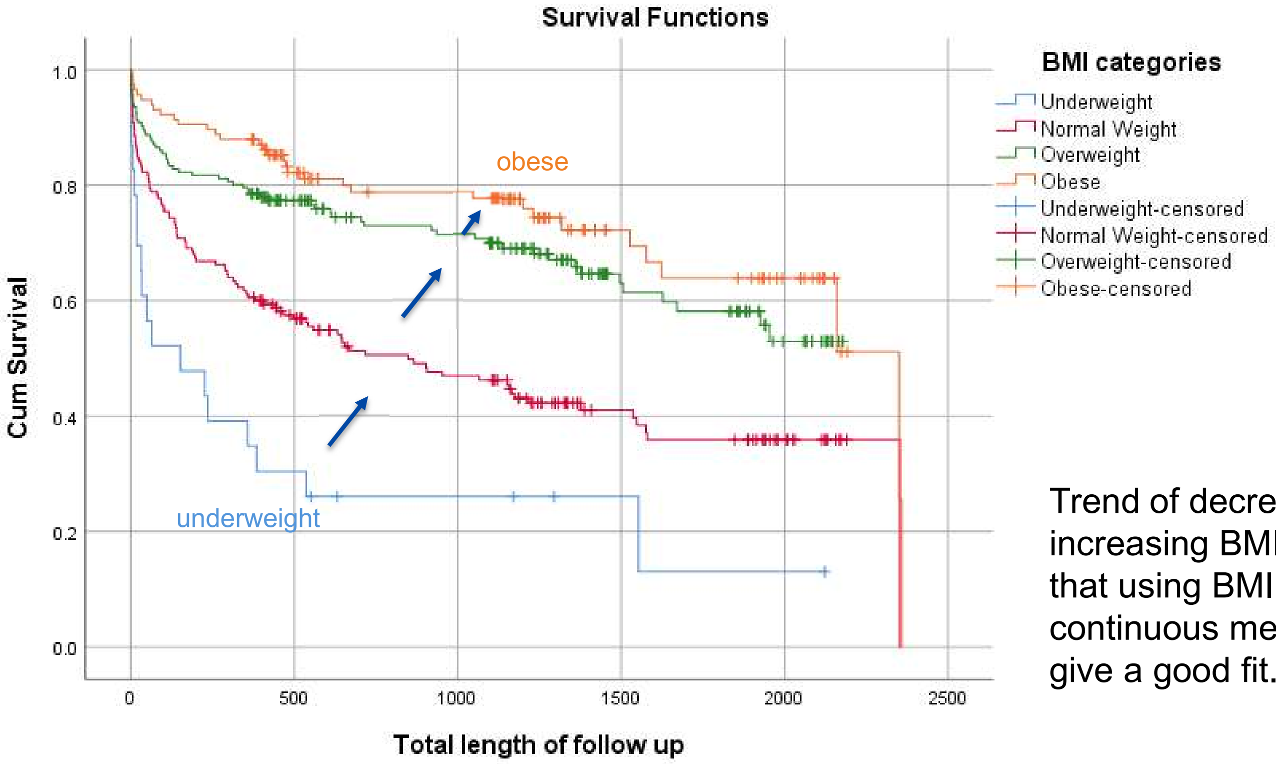
Question: Would you expect a linear relationship between BMI and survival?

Have a look at the Kaplan-Meier survival curves with these categories.

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Linearity of continuous predictors (eg BMI)



Trend of decreasing HR with increasing BMI indicates that using BMI as a continuous measure would give a good fit.

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Linearity of continuous predictors (eg BMI)

Now put BMI categories into the model instead of the continuous BMI predictor. Lets see if it gives a better fit.

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Linearity of continuous predictors (eg BMI)

model no.	Variables in the model	-2 log L	AIC -2LogL+2q
10	Age+Gender+BMI+HR	2286.8	2294.8
11	Age+Gender+BMI_cat+HR	2282.4	2294.4

The switch to BMI categories has a trivially lower AIC. Either of these choices would be acceptable depending on how you want to look at BMI.

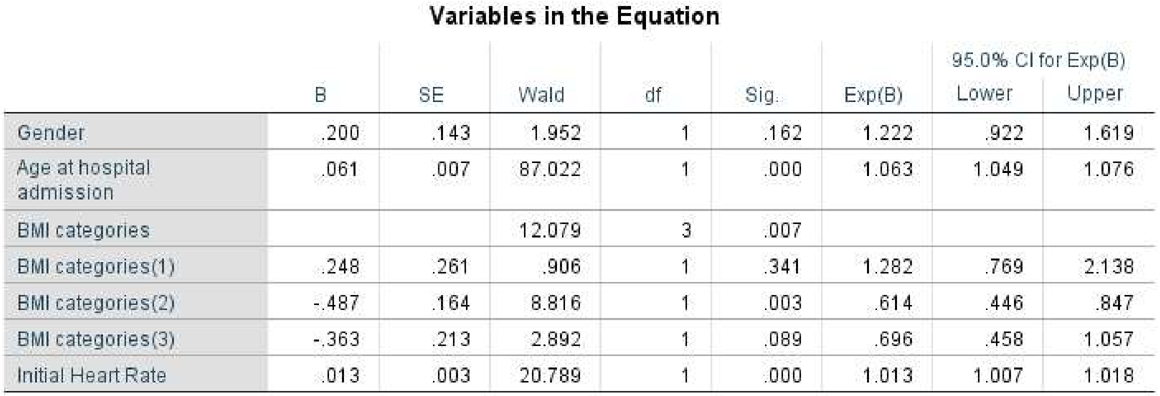
Another method for checking linearity of a continuous predictor is to plot the value of the predictor variable against the Martingale residuals for the null model. See Collett section 4.2.3

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Linearity of continuous predictors (eg BMI)

Model 11: Parameter Estimates



Variables in the Equation

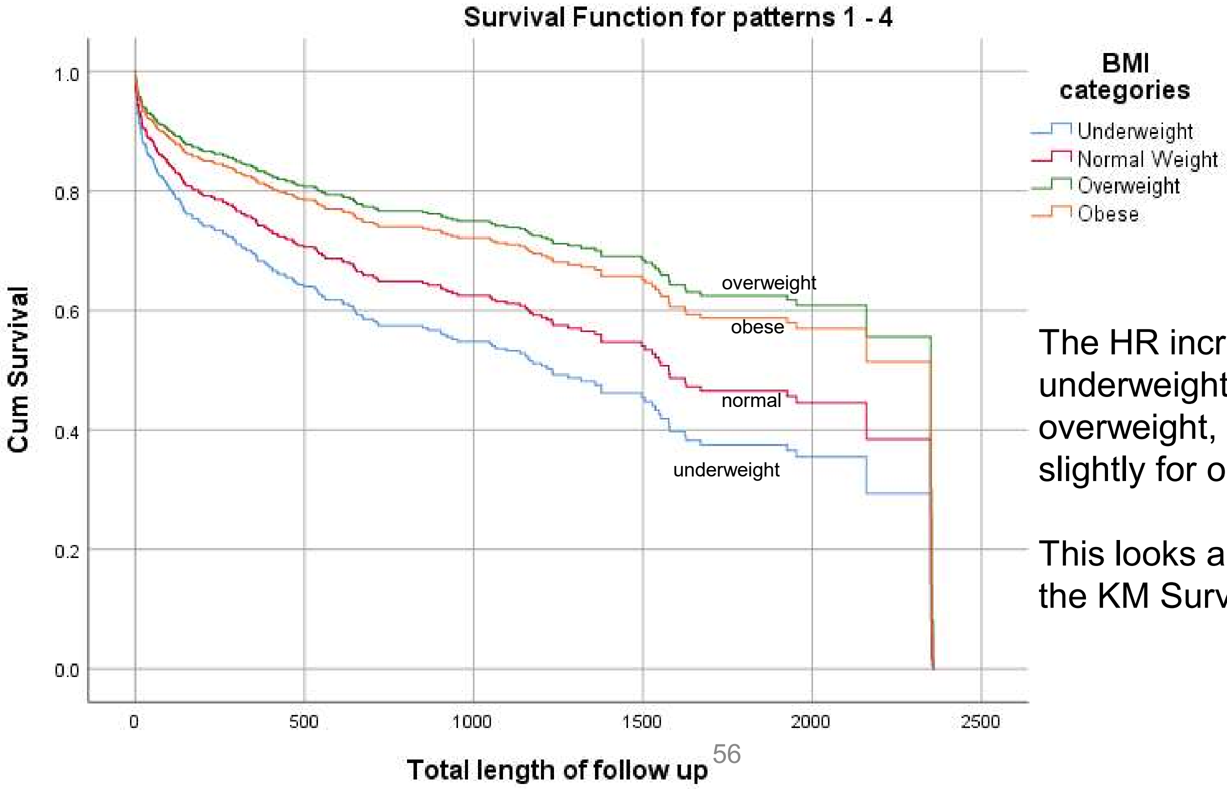
		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Male	Gender	.200	.143	1.952	1	.162	1.222	.922	1.619
	Age at hospital admission	.061	.007	87.022	1	.000	1.063	1.049	1.076
	BMI categories			12.079	3	.007			
underweight	BMI categories(1)	.248	.261	.906	1	.341	1.282	.769	2.138
overweight	BMI categories(2)	-.487	.164	8.816	1	.003	.614	.446	.847
Obese	BMI categories(3)	-.363	.213	2.892	1	.089	.696	.458	1.057
	Initial Heart Rate	.013	.003	20.789	1	.000	1.013	1.007	1.018

The table shows HR estimates for BMI categories compared to “normal weight” as the reference category. Other choices of reference category may be preferred.

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Linearity of continuous predictors (eg BMI)



The HR increases from underweight to normal to overweight, then decreases slightly for obese category.

This looks a bit different to the KM Survival curve

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Interactions between predictors

You should also investigate whether any significant interactions exist between predictors. In this example we should consider interactions between

- Age
- Gender
- BMI (category)
- Initial heart rate

Refer to analysis in Hosmer “Applied Survival Analysis” chapter 4.4 for a detailed description of how to investigate covariate confounders and interactions (also called ‘effect modifiers’) using this example.

Desired interaction terms would be added to the model in a similar fashion to linear or logistic regression.

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Fit the univariable models for Age and Gender, then the main effects model, then include the interaction term. See what happens to the HR estimates and p values.

Model	Variable	beta coeff	Std Err	p value (WALD)	HR
Age	Age	0.066	0.006	<0.001	1.068
Gender	Gender	0.381	0.138	0.006	1.464
A + G	Age	0.067	0.006	<0.001	1.069
	Gender	0.066	0.141	0.641	1.068
A + G + A*G	Age	0.048	0.010	<0.001	1.049
	Gender	-2.329	0.992	0.019	0.097
	Age*Gender	0.030	0.013	0.015	1.031

Gender becomes non-significant – why?

Cox Proportional Hazards Regression Workflow

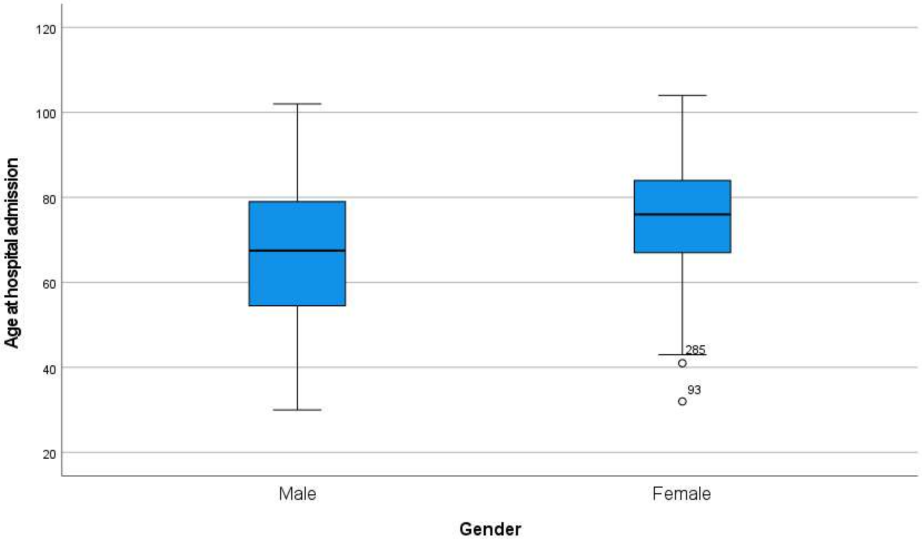
3. Model Fitting

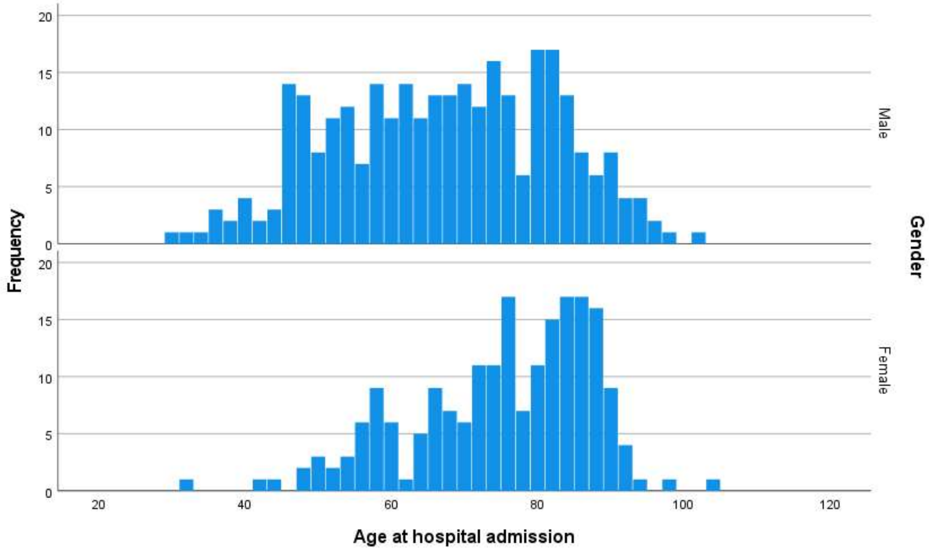
This is where some EDA will help.

Median age of males = 67.5

Median age of females = 76.0

If Age is important, then the effect of Gender in the simple model will be confounded by this effect.





Cox Proportional Hazards Regression Workflow

3. Model Fitting

The interaction term is significant – age is modifying the effect of gender.

We should include the interaction term in the model.

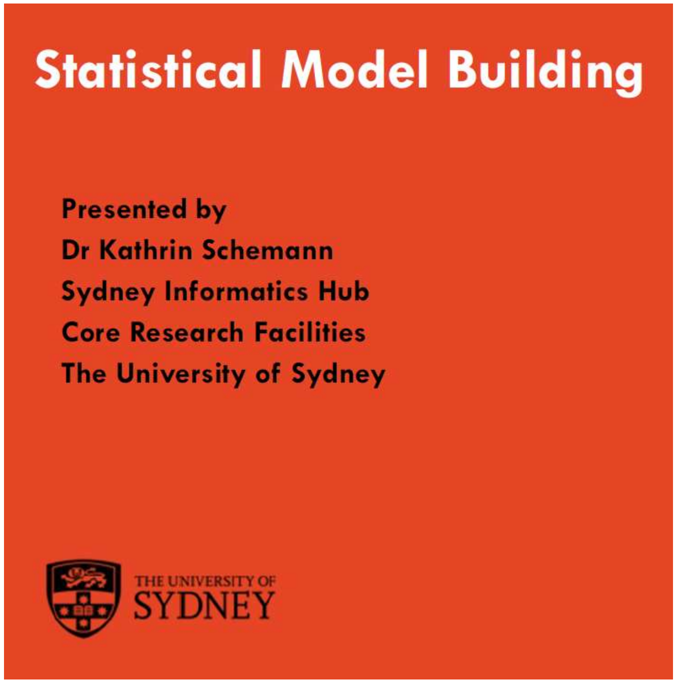
Model	Variable	beta coeff	Std Err	p value (WALD)	HR
Age	Age	0.066	0.006	<0.001	1.068
Gender	Gender	0.381	0.138	0.006	1.464
A + G	Age	0.067	0.006	<0.001	1.069
	Gender	0.066	0.141	0.641	1.068
A + G + A*G	Age	0.048	0.010	<0.001	1.049
	Gender	-2.329	0.992	0.019	0.097
	Age*Gender	0.030	0.013	0.015	1.031

Cox Proportional Hazards Regression Workflow

3. Model Fitting

Other training and resources:

Attend “Statistical Model Building” workshop for a more complete overview of this topic.



Statistical Model Building

Presented by
Dr Kathrin Schemann
Sydney Informatics Hub
Core Research Facilities
The University of Sydney

Cox Proportional Hazards Regression Workflow

4. Check the model assumptions

- Proportional hazards assumption
- Leverage and influence (outliers)
- Goodness of fit

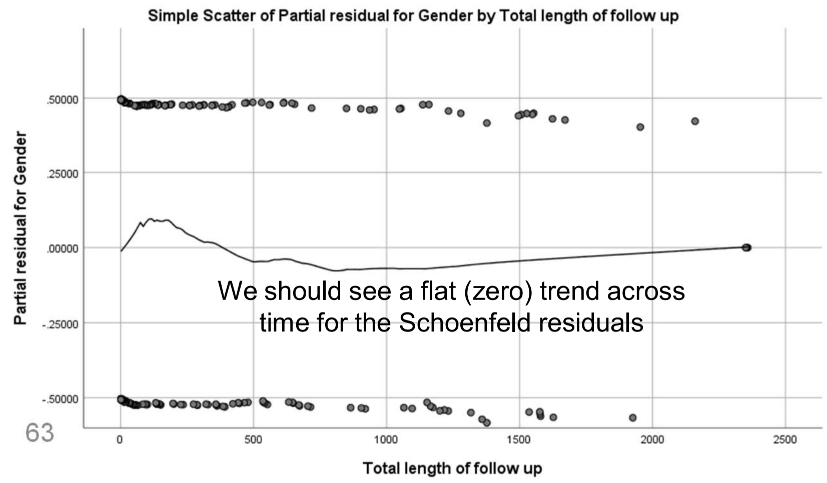
Cox Proportional Hazards Regression Workflow

4. Check the model assumptions

Proportional hazards assumption: using residuals

- Many types of residuals exist (Schoenfeld, Cox-Snell, Deviance, Martingale, etc) and interpretation varies.
- Many residual plots exhibit patterns even when the model is correctly fitted!
- Interpretation of residuals is not as easy as with linear regression.
- Many factors need to be taken into account.
- See references for further information.

Plot of partial (Schoenfeld) residuals for Gender against time



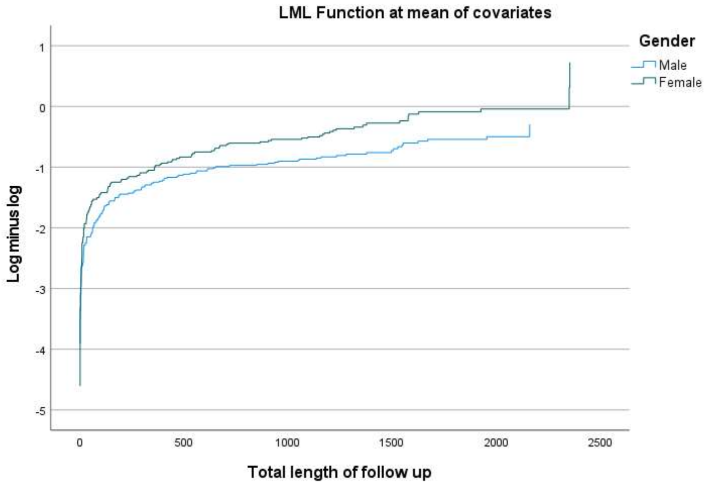
We should see a flat (zero) trend across time for the Schoenfeld residuals

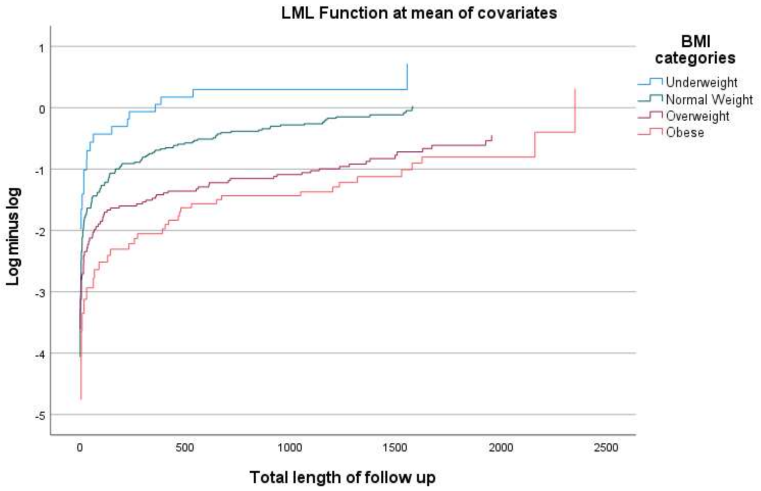
Cox Proportional Hazards Regression Workflow

4. Check the model assumptions

Proportional hazards assumption: using Graphical methods

Look at the LML curves





Gender looks OK. The BMI categories are not always proportional. Note the obese line wanders around especially near the end of the study period (when uncertainty is higher).

Cox Proportional Hazards Regression Workflow

4. Check the model assumptions

Proportional hazards assumption: using Time-dependent covariates

- We can add interaction terms into the Cox regression that include the “time” variable. This is **“Cox Regression with time-dependent covariates”**
- For our chosen model we will separately test the following interaction terms:
  - Time*Gender
  - Time*Age
  - Time*BMI_cat
  - Time*HR

- These will be added to the full model (so four models to check)

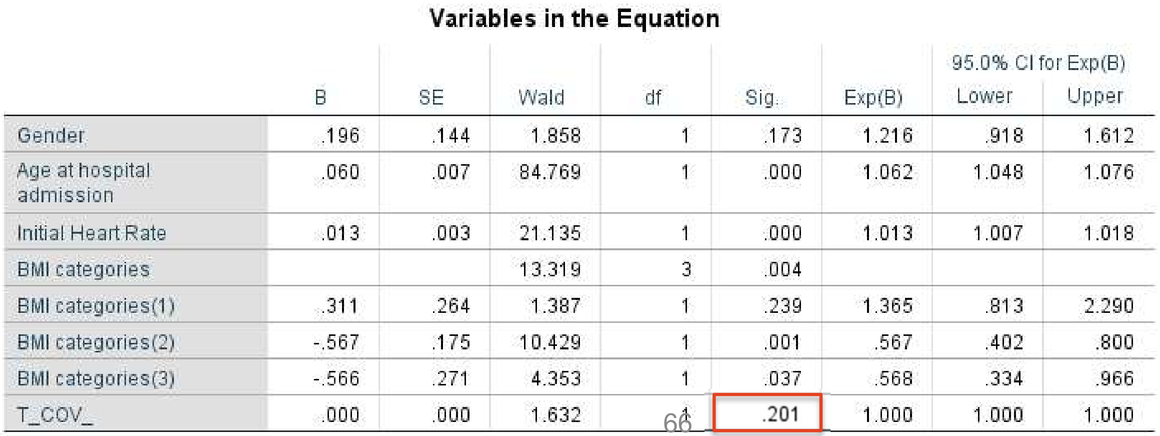
Cox Proportional Hazards Regression Workflow

4. Check the model assumptions

Proportional hazards assumption: using Time-dependent covariates

Check the significance of the “T_Cov_” time interaction terms.

The model including Time*BMI_cat is shown below. The interaction term is not significant so we can say that BMI_cat is time independent. (Not shown: The interaction terms for the other 3 variables were also not significant.)



Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Gender	.196	.144	1.858	1	.173	1.216	.918	1.612
Age at hospital admission	.060	.007	84.769	1	.000	1.062	1.048	1.076
Initial Heart Rate	.013	.003	21.135	1	.000	1.013	1.007	1.018
BMI categories			13.319	3	.004			
BMI categories(1)	.311	.264	1.387	1	.239	1.365	.813	2.290
BMI categories(2)	-.567	.175	10.429	1	.001	.567	.402	.800
BMI categories(3)	-.566	.271	4.353	1	.037	.568	.334	.966
T_COV_	.000	.000	1.632	1	.201	1.000	1.000	1.000

Cox Proportional Hazards Regression Workflow

4. Check the model assumptions

Leverage and influence (outliers)

- There are different techniques for identifying influential and poorly fit values – in a similar fashion to those used in linear regression.
- Option 1: scaled score residuals
- Option 2: likelihood displacement vs Martingale residuals
  (see Hosmer Lemeshow and May “Applied Survival Analysis” for further details)

Cox Proportional Hazards Regression Workflow

4. Check the model assumptions

Goodness of Fit

- Can compare the observed and expected events (across G groups where G=integer(no. of events/40)
  [refer to Hosmer and Lemeshow "Applied Survival Analysis" for details]
- "Pseudo" measures analogous to R^2 found in linear regression have been proposed by Nagelkerke (1991), O'Quigley (2005) and Royston (2006).

Cox Proportional Hazards Regression Workflow

4. Check the model assumptions

Goodness of Fit

- "Pseudo" R^2 by Nagelkerke (1991)

$$R_p^2 = 1 - \left\{exp\left[\frac{2}{n}\left(L_0 - L_p\right)\right]\right\}$$

Where:

L_p = log partial likelihood for the fitted model with p covariates

L_0 = log partial likelihood for the null model

n = number of events

Availability of goodness of fit statistics will vary by software. Pseudo R^2 is given in survival::coxph in R, but not in SPSS.

Cox Proportional Hazards Regression Workflow

5. Interpret the model

Hazard Ratios

- HR’s are similar to Odds Ratios, but express a comparative measure (a rate) over the entire study period.
- The Hazard Ratio can be interpreted as a predicted change in the hazard for a unit increase in the predictor.
- HR’s for continuous predictors should be expressed in clinically relevant units. For example if age is a covariate, we could report the HR per year change, or the HR per decade change. For some covariates the HR per standard deviation change might be useful.

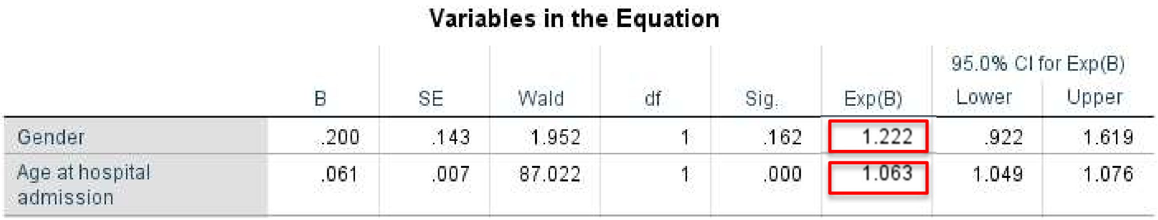
Cox Proportional Hazards Regression Workflow

5. Interpret the model

Hazard Ratios – from WHAS example

Example of reporting language:

- The mortality hazard for females is 1.2 times [95% CI: 0.92-1.62] that of males.
- The mortality hazard is increased by 6.3% [95% CI: 4.9-7.6%] for each additional year of age of the patient.
- Note: if the interaction of Age*Gender is found to be significant we should report on how these interact.



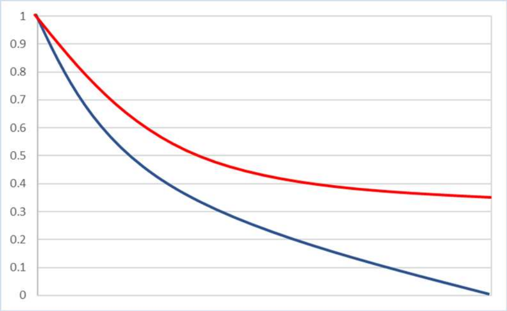
Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Gender	.200	.143	1.952	1	.162	1.222	.922	1.619
Age at hospital admission	.061	.007	87.022	1	.000	1.063	1.049	1.076

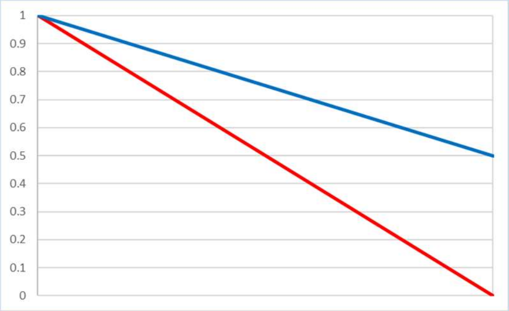
Cox Proportional Hazards Regression
Challenge questions

Q1: Which of these survival curves indicates that the data does not meet the assumption of proportional hazards?

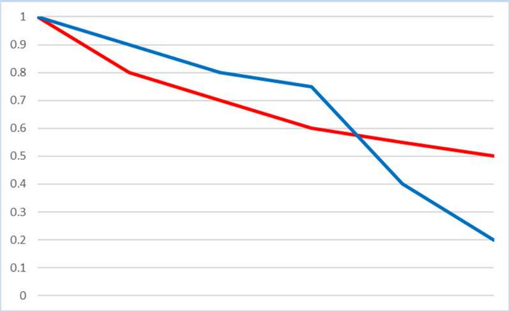
(a)



(b)



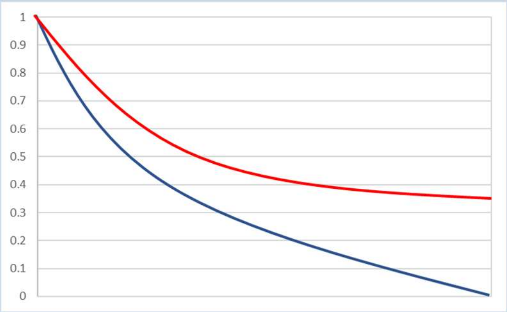
(c)



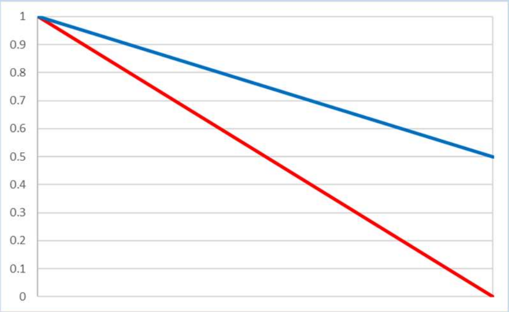
Cox Proportional Hazards Regression
Challenge questions

Q1: Which of these survival curves indicates that the data does not meet the assumption of proportional hazards?

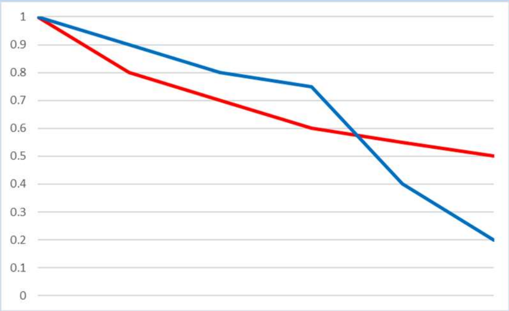
(a)



(b)



(c)



Cox Proportional Hazards Regression
Challenge questions

Q2: What is Sir David Cox's middle name?

a) Box

b) Roxbee

c) Furminster

d) All of the above

Cox Proportional Hazards Regression Challenge questions

Q2: What is Sir David Cox's middle name?

a) Box

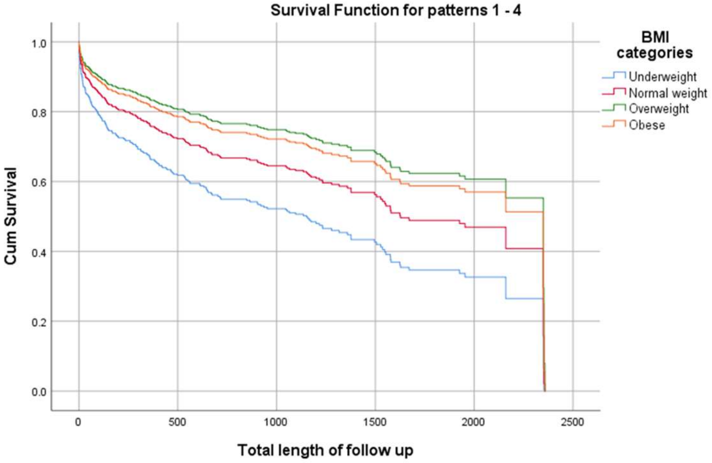
[✓ b) Roxbee]

c) Furminster

d) All of the above

Cox Proportional Hazards Regression

Any questions?



Cox, Box or Fox?

Survival Analysis other models

3. Parametric regression models – like Cox, but assumes an underlying survival distribution like exponential or Weibull.

This is useful for prediction as these models make strong assumptions about the rate of survival over time

Survival Analysis other models

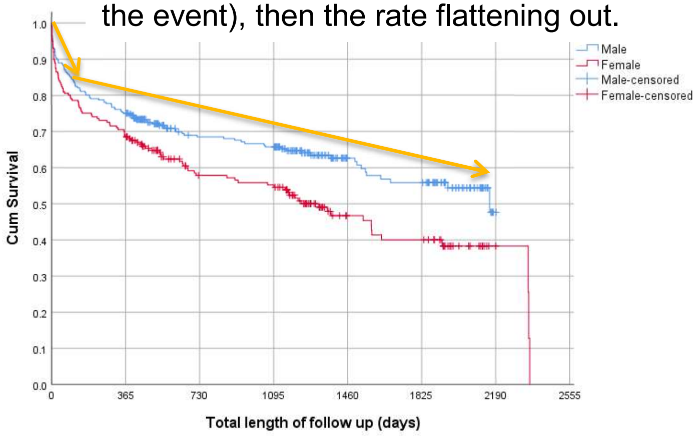
4. Frailty models

- Takes account of heterogeneity of subjects in relation to the event occurring using a “random intercept” like in Mixed Models

Clustering or Shared Frailty (or just random effects model)

- Multiple events for the same person
- Multiple sites on the same person

Frailty is often observed as a high hazard rate early on (when frail individuals suffer the event), then the rate flattening out.



Frailty models can be difficult to implement
An alternative is to use a “stratified” model when cluster sizes are large.

Survival Analysis other models

5. Cox Regression with time varying covariates

- Can be used to check proportional hazards assumption
- Can be used when covariate HR changes with time
- Can be used when the value of the covariate changes with time

References - Software

Software	accessibility	features
SPSS	Available to USyd staff and students	Kaplan Meier Cox Regression
STATA	via subscription	Kaplan Meier Cox Regression
R packages: survival, survminer, survPen	free open source	K-M, Cox Regression Huge variety of options in these and other packages
SAS	Some availability for USyd staff and students	KM and Cox proc phreg, lifetest, lifereg
GraphPad PRISM	Some availability for USyd staff and students	Kaplan Meier and Cox
MedCalc	via subscription (annual or lifetime)	Kaplan Meier Cox Regression

References

Survival Analysis examples

Paper citation and link	Comment
Altinbas, M et al. "A Randomized Clinical Trial of Combination Chemotherapy with and Without Low-molecular-weight Heparin in Small Cell Lung Cancer." Journal of thrombosis and haemostasis 2.8 (2004): 1266–1271. Web.	Examples of reporting Kaplan-Meier figures
Abe, Tetsuya et al. "Randomized Phase III Trial Comparing Weekly Docetaxel Plus Cisplatin Versus Docetaxel Monotherapy Every 3 Weeks in Elderly Patients with Advanced Non-Small-Cell Lung Cancer: The Intergroup Trial JCOG0803/WJOG4307L." Journal of clinical oncology 33.6 (2015): 575–581. Web.	Shows Cox HR's on a Forest Plot style figure
Batson, Sarah et al. "Review of the Reporting of Survival Analyses Within Randomised Controlled Trials and the Implications for Meta-Analysis." PloS one 11.5 (2016): e0154870–e0154870. Web.	Advice on reporting Survival Analysis

References

VIDEOS

Marinstats lectures https://youtu.be/vX3l36ptrTU

WEBSITES

UCLA IDRE https://stats.idre.ucla.edu/r/dae/mixed-effects-cox-regression/ has example R code

The Analysis Factor https://www.theanalysisfactor.com/resources/by-topic/survival-analysis/

BOOKS

Collett, David. Modelling Survival Data in Medical Research, Third Edition. CRC Press, 2015. Print. https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/1367smt/scopus2-s2.0-85053657101

Hosmer, David W. Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition. Wiley Blackwell, 2011. Web. https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/1367smt/scopus2-s2.0-84947789021

Moore, Dirk F. Applied Survival Analysis Using R. Cham: Springer International Publishing, 2016. Print. https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/14vvljs/alma991014234299705106

Pintilie, Melania. Competing Risks : a Practical Perspective . Chichester, England ;: John Wiley & Sons, 2006. Print. https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/14vvljs/alma991010555469705106

Further Assistance: Sydney University

SIH

– **Personal Consultations** can be requested via our website:

www.sydney.edu.au/research/facilities/sydney-informatics-hub.html OR Google “Sydney Informatics Hub”

– **Training** Sign up to our mailing list to be notified of upcoming training: https://signup.e2ma.net/signup/1945889/1928048/

– **Hacky Hour** www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training/hacky-hour.html OR Google “Sydney Hacky Hour”

OTHER

– **Open Learning Environment (OLE) courses**

– **Linkedin Learning:** https://linkedin.com/learning/

Acknowledging SIH

All University of Sydney resources are available to Sydney researchers **free of charge.** The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording:

General acknowledgement:

"The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

Acknowledging specific staff:

“The authors acknowledge the technical assistance of (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney.”

For further information about acknowledging the Sydney Informatics Hub, please contact us at **sih.info@sydney.edu.au**.

End of Workshop on Survival Analysis

- Thank you for your interest and attention
- Questions and comments welcome
- We appreciate your feedback via the on-line survey

Kathrin Schemann BAnVetBioSc (Hons) MBiostat PhD
Senior Consultant: Statistics
The University of Sydney
Sydney Informatics Hub | Core Research Facilities
Merewether Building (H04) | The University of Sydney |
NSW | 2006
+61 422 063 370
Kathrin.schemann@sydney.edu.au | sydney.edu.au