Introduction to Survival Analysis

Presented by Dr Kathrin Schemann
Senior Statistical Consultant
Sydney Informatics Hub
Core Research Facilities
sydney.edu.au/sydney-informatics-hub

THE UNIVERSITY OF SYDNEY

Slides available here

Acknowledging SIH

- All University of Sydney resources are available to 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 for use of workshops and workflows:

- *“The authors acknowledge the Statistical workshops and workflows provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney.”*

What is a workflow?

- Every statistical analysis is different, but all follow similar paths. It can be useful to know what these paths are.

- We have developed practical, step-by-step instructions that we call ‘workflows’, that can you can follow and apply to your research.

- We have a general research workflow that you can follow from hypothesis generation to publication.

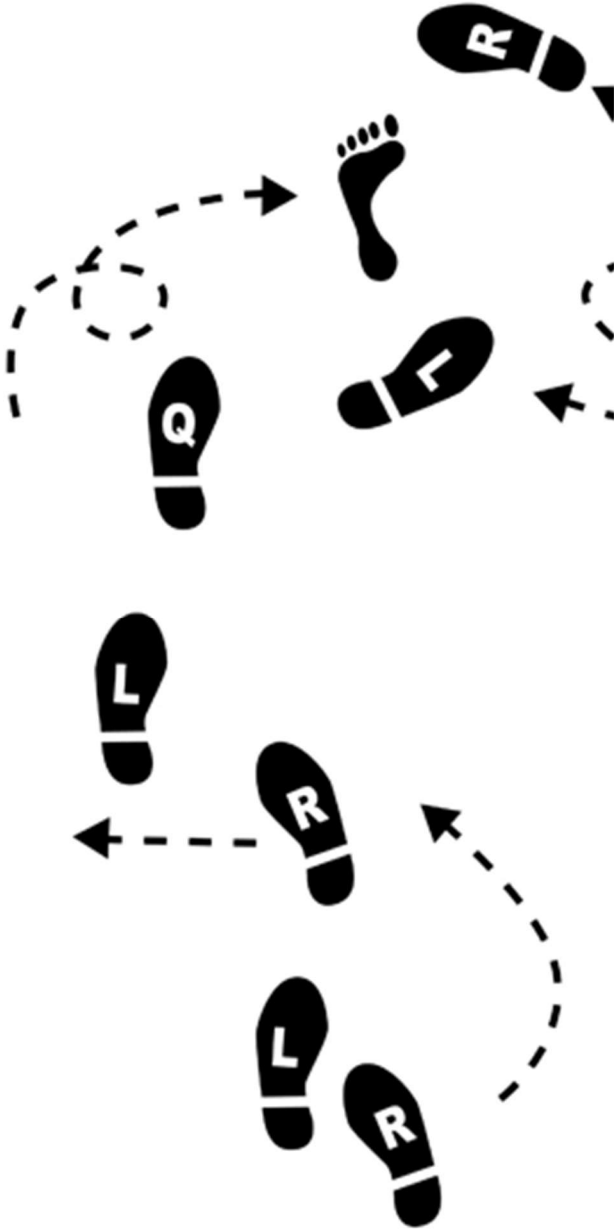
- And statistical workflows that focus on each major step along the way (e.g. experimental design, power calculation, model building, analysis using linear models/survival/multivariate/survey methods).

Statistical workflows

- Our statistical workflows can be found within our workshop slides.

- Statistical workflows are software agnostic, in that they can be applied using any statistical software.

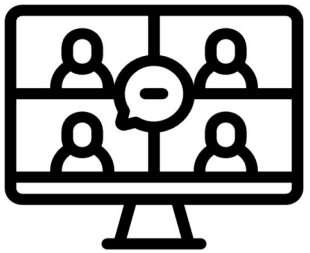
- To access these statistical workflows and more, visit our Workshops and Workflows page.



Software workflows

- There may also be accompanying software workflows that show you how to perform the statistical workflow using particular software packages (e.g. R or SPSS). We won't be going through these in detail during the workshop. If you are having trouble using them, we suggest you attend our monthly Hacky Hour where SIH staff can help you.
- Our software workflows contain:
    - R code and comments.
    - SPSS syntax as well as screenshots of the point and click procedures and written methods.
    - Screenshots of the point and click procedures and written methods for other bespoke software.

How to use our workshops



Workshops developed by the Statistical Consulting Team within the Sydney Informatics Hub form an integrated modular framework. Researchers are encouraged to choose modules to create custom programs tailored to their specific needs. This is achieved through:

- Short 90-minute workshops, acknowledging researchers rarely have time for long multi day workshops.
- Providing statistical workflows appliable in any software, that give practical step by step instructions which researchers return to when analysing and interpreting their data or designing their study e.g. workflows for designing studies for strong causal inference, model diagnostics, interpretation and presentation of results.
- Each one focusing on a specific statistical method while also integrating and referencing the others to give a holistic understanding of how data can be transformed into knowledge from a statistical perspective from hypothesis generation to publication.

For other workshops that fit into this integrated framework, refer to our training link page under statistics, found at Workshops and training

During the workshop

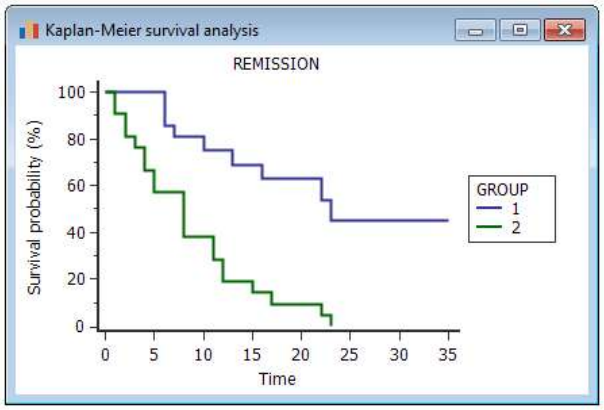
- Ask short questions or clarifications during the workshop (either by Zoom chat or verbally). There will be breaks during the workshop for longer questions.

- Slides with this blackboard icon are mainly for your reference, and the material will not be discussed during the workshop.

- Challenge questions will be encountered throughout the workshop.

Workshop Aims

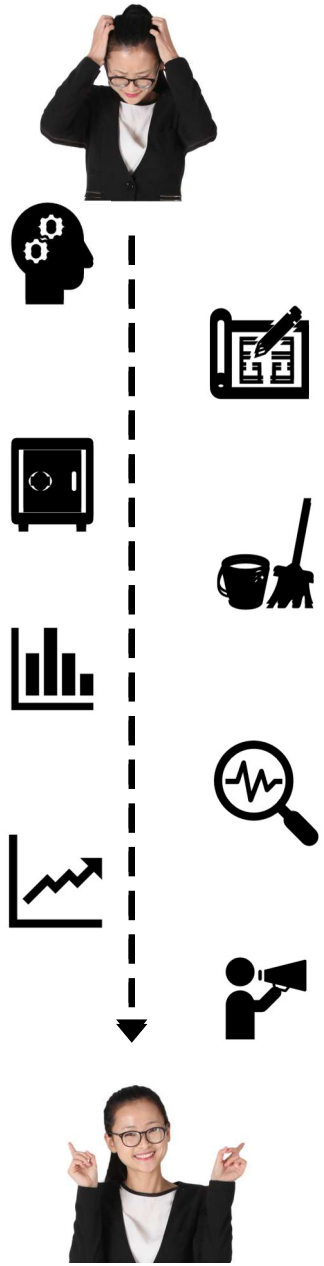
- Understand the key concepts in Survival Analysis



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

General Research Workflow

1. **Hypothesis Generation** (Research/Desktop Review)
2. **Experimental and Analytical Design** (Sampling, power, ethics)
3. **Collect/Store Data**
4. **Data cleaning**
5. **Exploratory Data Analysis** (EDA)
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. **Publication**



Workshop contents

General Research Workflow:

1. **Hypothesis Generation**
2. **Experimental and Analytical Design**
3. **Collect/Store Data**
4. **Data cleaning**
5. **Exploratory Data Analysis** (EDA)
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. **Reporting for publication**

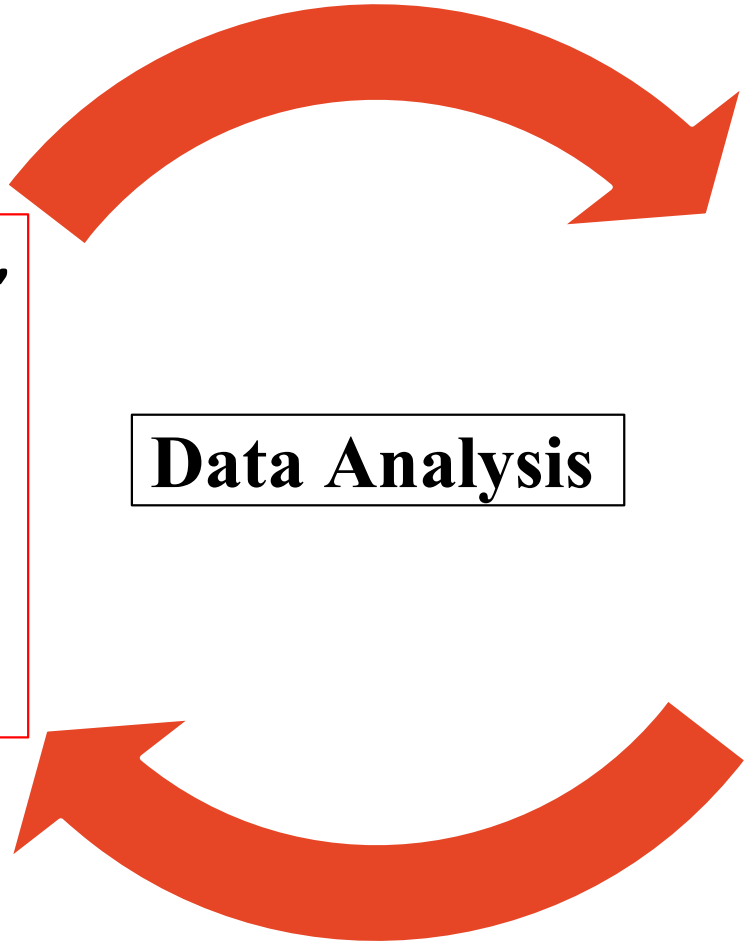
Steps in Survival Analysis:

1. **Study and Analytical Design**
2. **Data cleaning**
3. **Exploratory Data Analysis** (EDA) – Kaplan Meier Plot
4. **Data Analysis aka inferential analysis –** Kaplan Meier Test (non-parametric); Cox PH regression model (semi-parametric)
5. **Reporting for publication**

Data Analysis step

Fit a single model- LM1/2/3, Survival analysis

1. Check model assumptions
2. Check goodness-of-fit
3. Interpreting Model Parameters and reach a conclusion



Iteratively fit models – see our Model Building Workshop

- Pick predictors to fit and a suitable modelling method based on EDA
- Iteratively fit and assess models using predetermined strategy until no more criteria are met.

Introduction

When to use Survival Analysis?

- Outcome = time elapsed until a specified event occurs
  - Can be used for different study types, e.g. in a Randomised Controlled Trial (RCT), observational study (e.g. cohort study, case-control study) or reliability or environmental/biological monitoring study
  - The classic event is “death” which gives survival analysis its name.
  - The event doesn’t have to occur for all subjects. This is an important feature of survival analysis – censoring.

Alternative analysis:
Logistic regression (Generalised Linear Model- GLM) - models the probability of the binary event occurring within a timeframe, not the rate over time.

Survival Time and Event

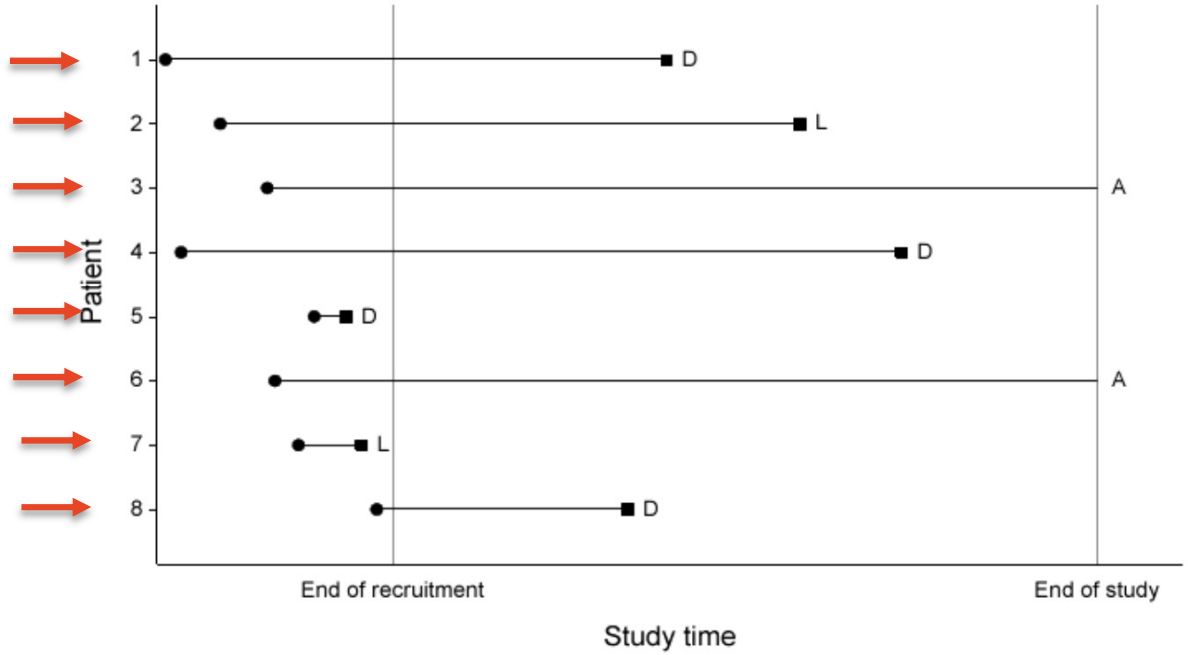
Examples

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

Define the 'time-to-event', i.e. the time beginning and end points.

Define event – data type must be binary.

Sampling + different types of observations in Survival Analysis



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

Survival Analysis - 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.

Survival Analysis *assumptions* around censoring that need to be checked:

1. Verify that the censoring is non-informative; that is, the mechanism causing the censoring is independent of the event mechanism – assess using sensitivity analysis for different types of censoring.

2. In cases where the censoring might be informative, consider **joint modeling techniques-** Joint modeling in presence of informative censoring on the retrospective time scale with application to palliative care research - PMC.

The three types of censoring in Survival Analysis

Censoring is a type of data structure that needs to be considered when choosing the type of analysis to use to avoid bias.

Sensitivity analysis to assess the impact of different censoring mechanisms on the final outcome may be useful.

1. ***Left Censoring***: Occurs when the event has already happened before the start of observation.

Examples:
Early detection, e.g. biomarker or environmental studies where tests measurements may be below detectable level or have already reached a set event threshold at time of first observation.

Analysis implications:
Statistical imputation techniques may be used to estimate event time + analysis methods for survival data that account for censoring.

2. ***Right Censoring***: Occurs when the event has not happened by the end of the observation period – most typical.

Examples: Longitudinal, clinical studies – the methods presented in this workshop account for censoring and are suitable.

3. ***Interval Censoring***: Occurs when the exact timing of the event is unknown but falls within a known interval.

Examples: Machine only checked periodically for reliability; or people screened periodically in medical studies.

Analysis implications: Requires more advance methods such as Turnbull Estimator and parametric modelling methods

Workflow: Steps in Survival Analysis

1. **Experimental and Analytical Design**
2. **Data cleaning**
3. **Exploratory Data Analysis** (EDA)
   - Survival data descriptive statistics
   - Kaplan-Meier procedure - plot
4. **Data Analysis aka inferential analysis**
   - Kaplan Meier associated test (non-parametric)
   - Cox PH regression model (semi-parametric)
   - Advanced models (parametric) and other model types
5. **Reporting for publication**

Step 1: Study and Analytical Design – Survival Analysis

Tips from the Consulting Room:

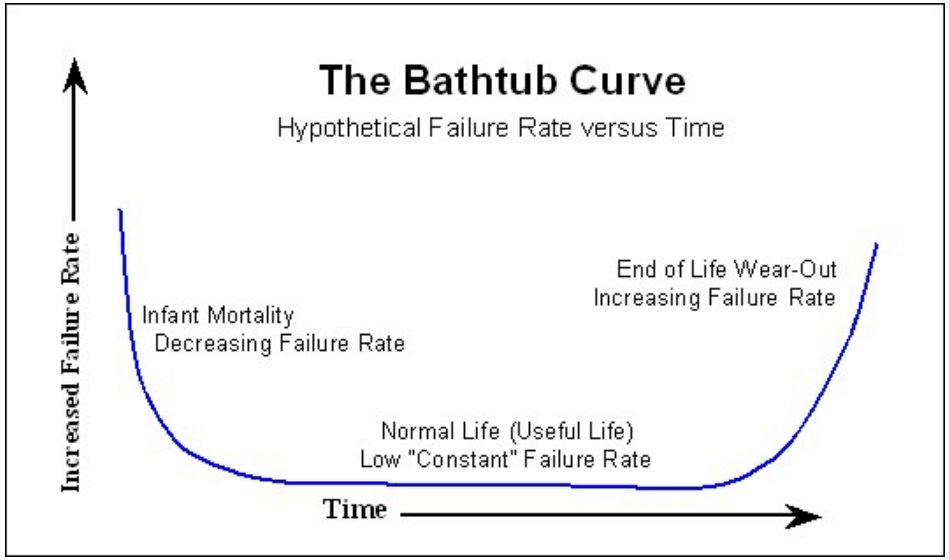
- Consider creating a Statististical Analysis Plan (SAP) - **Research Essentials** workshop
- For multiple predictors, specify your Model Building Strategy in your SAP – see our **Model Building** workshop
- Ensure **sufficient sample size and statistical power** to achieve meaningful results - in survival analysis this **depends on the number of events occurring**, NOT the number of samples/participants.

Take away: A sound study design and analysis plan increase your chances of grant success and high impact publications! Consider the SCU @SIH ***Experimental Design*** and ***Power + Sample Size*** workshops and consultations.

Step 1: Study and Analytical Design and Sampling

Consider what you expect to happen naturally over time over the course of your study period and how it could bias your analysis or make your chosen modelling approach unsuitable, e.g.:

- 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 PH model = semi-parametric model: Does not assume an underlying distribution of survival time. Dependence on time is unspecified.
- If recruiting young or old participants, you want to adjust analyses for time-varying confounding by Age.

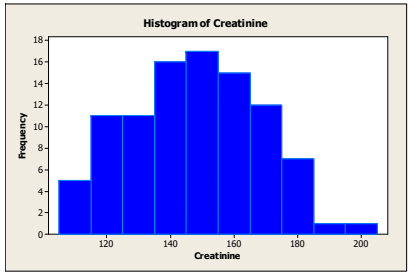


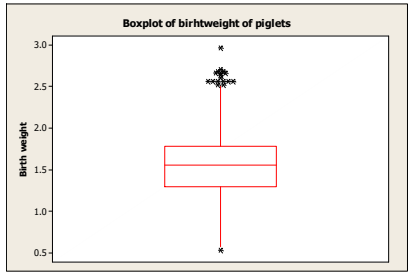
Step 2: Data Cleaning

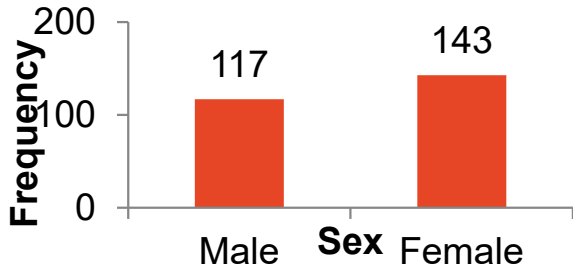
Consider a data dictionary.
For each variable: identify the data type (numeric/categorical) and use appropriate summary statistics and plotting to check the distribution:

- Numeric variable – histogram/boxplot; mean, median, standard deviation, percentiles, etc.
- Categorical variable – bar charts; frequency tables with count and percent

➔ See our Research Essentials – Analysing your Data Workshop for further details



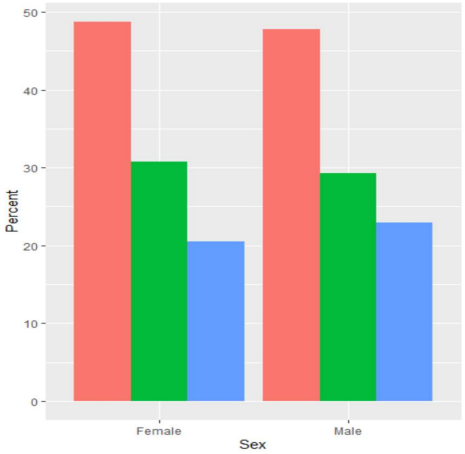




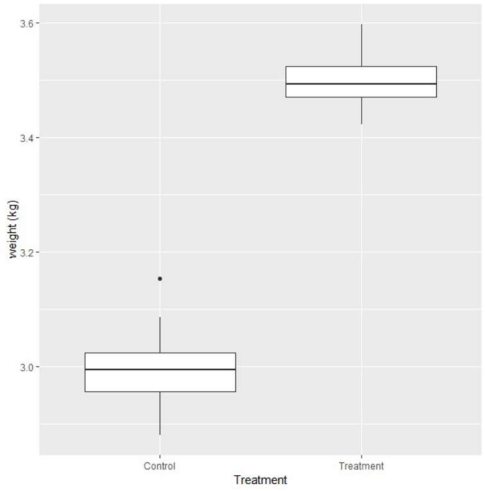
Step 3: Exploratory Data Analysis (EDA)

3.1 Assess relationships between each predictor and the outcome. Tabulate and plot:

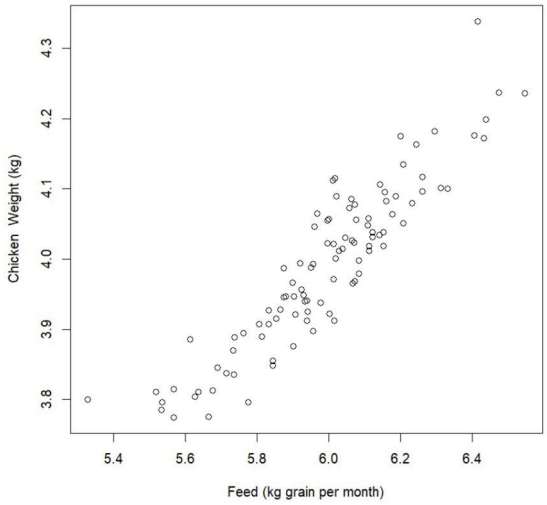
- Categorical variables: Count and percentage for cross-tabulation
- Numeric variables: Summary statistics by categories or Paersons correlation coefficient.
- Plot variables over time and the relationships– see ***Research Essentials*** workshop.



2 categorical variables
- side-by-side bar charts



1 categorical, 1 numeric variable
- side-by-side boxplots



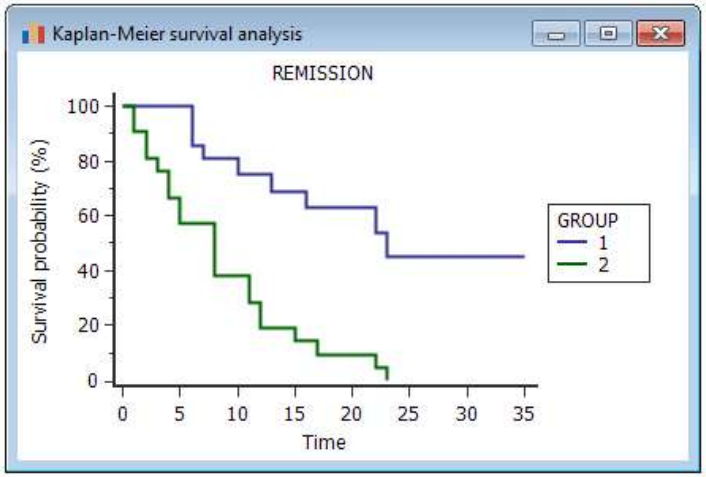
2 numeric variables
– xy scatter plot

3. EDA for Survival Analysis

Assess relationships between each predictor and the survival time outcome + event occurrence and censoring by groups of categorical variables. Plot data over time – plot survival curves.

Case Processing Summary

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



Workflow: Steps in Survival Analysis

1. **Study and Analytical Design**
2. **Data cleaning**
3. **Exploratory Data Analysis** (EDA)
   - Kaplan Meier procedure for survival descriptive statistics and plots
4. **Data Analysis aka inferential analysis**
   - Kaplan-Meier univariable tests, e.g. Log-rank test (non-parametric)
   - Cox PH regression model (semi-parametric; multivariable modelling possible)
   - Advanced models (parametric) + other types of models
5. **Reporting for publication**

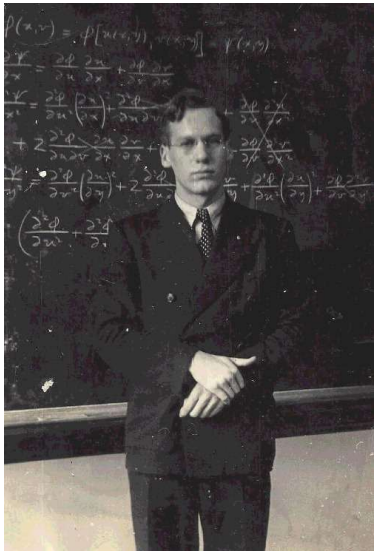
Kaplan-Meier Introduction

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.

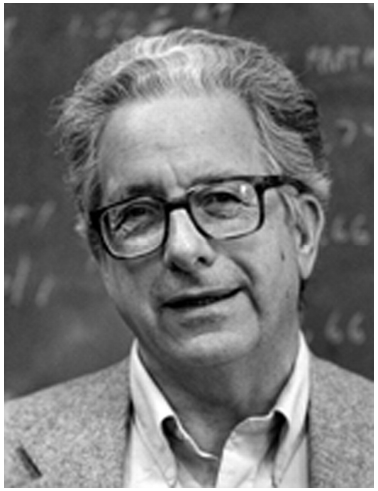
Their paper is the eleventh most cited scientific paper of the modern era (@ 2014).



Kaplan

-

Meier

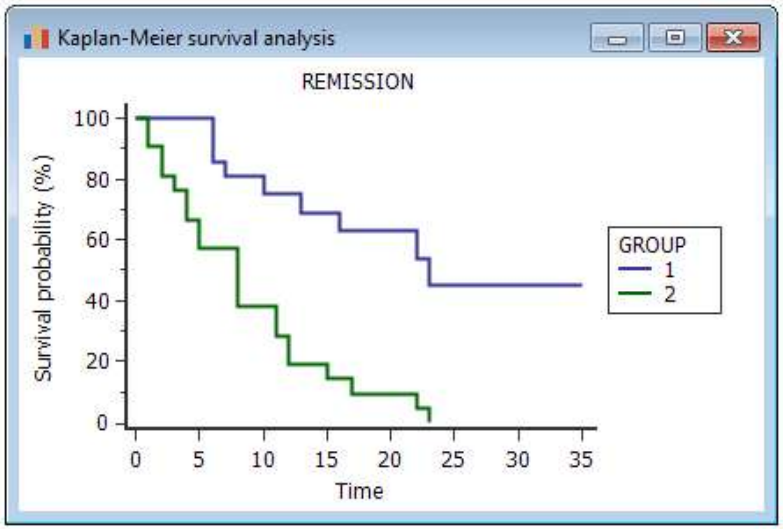


Kaplan-Meier – What is it all about?

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 - I know, an equation… bear with me please!

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 – Survival table

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

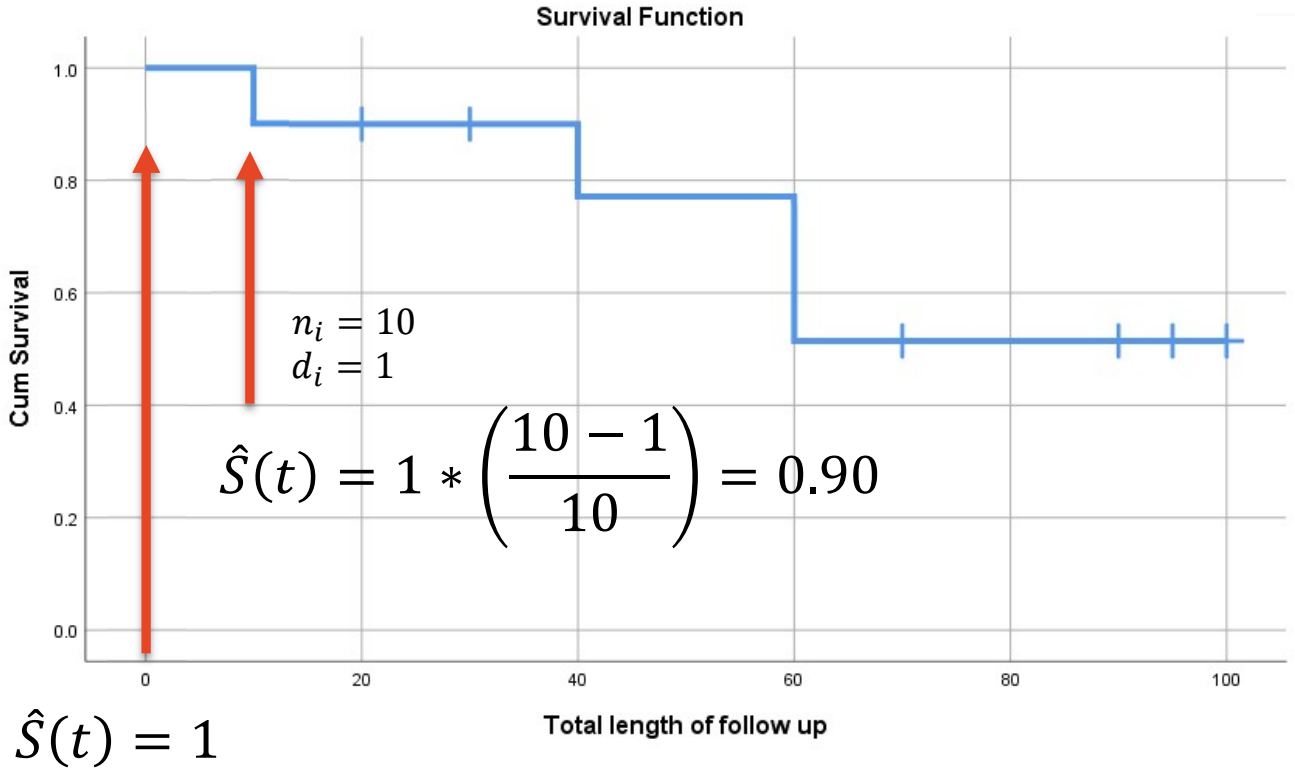
	Time	Status	Cumulative Proportion Surviving at the Time: Estimate	Cumulative Proportion Surviving at the Time: Std. Error	N of Cumulative Events	N of Remaining Cases
1	10.000	Dead	.900	.095	1	9
2	20.000	Alive	.	.	1	8
3	30.000	Alive	.	.	1	7
4	40.000	Dead	.771	.144	2	6
5	60.000	Dead	.	.	3	5
6	60.000	Dead	.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 – Survival curve

Illustration of Survival function: example

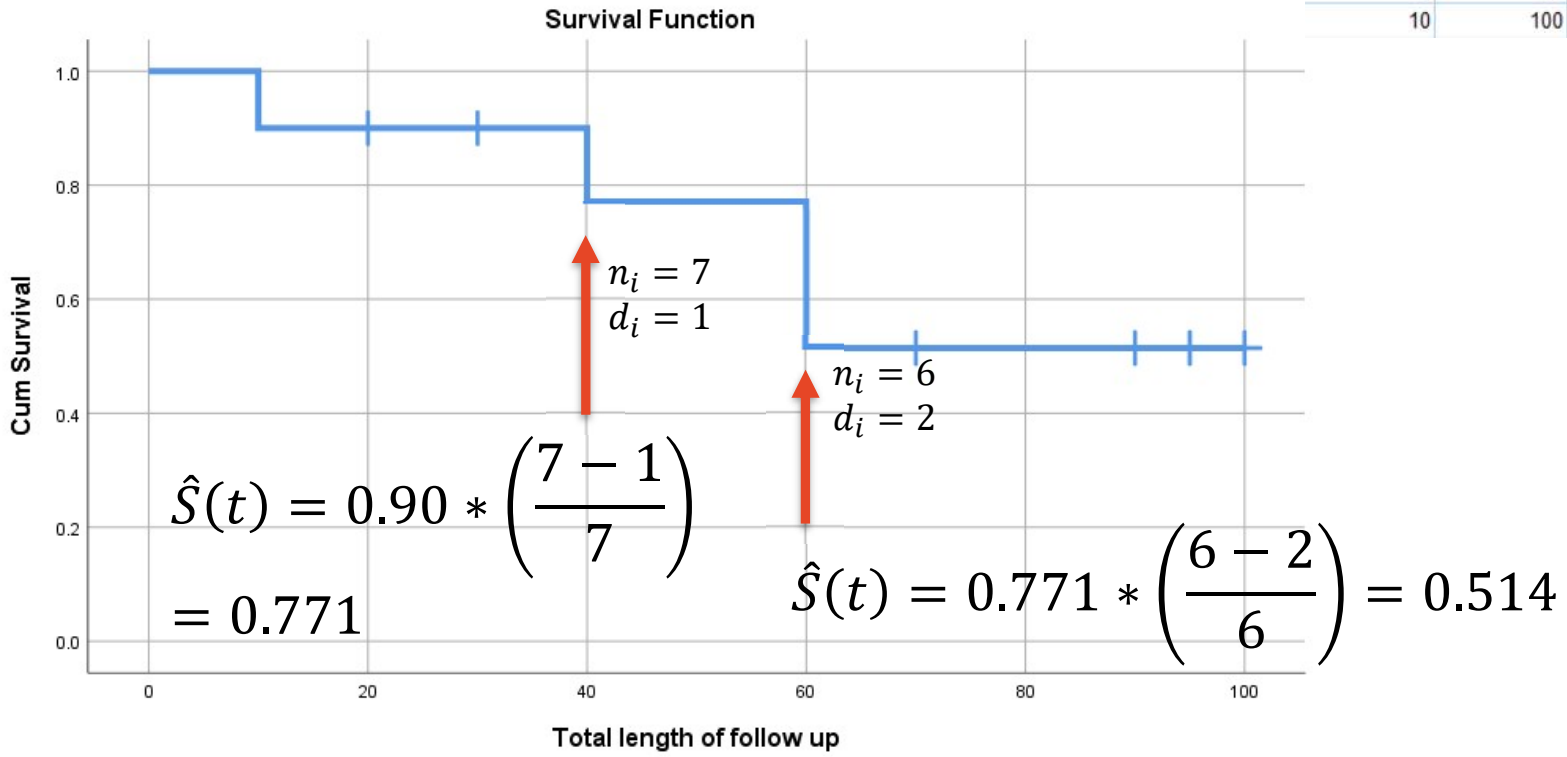
ID	lenfol	fstat
1	10	Dead
2	20	Alive
3	30	Alive
4	40	Dead
5	60	Dead
6	60	Dead
7	70	Alive
8	90	Alive
9	95	Alive
10	100	Alive



Kaplan-Meier - Survival curve

Illustration of Survival function: example

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 procedure - Workflow

With the Kaplan-Meier procedure we can plot the survival curves for an event and compare a single categorical predictor variable (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 univariable test statistics such as a Log-rank test.
3. Interpretation - Interpret the results

Kaplan-Meier Workflow - Step 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)

Explanatory variable/ categorical predictor of interest: Sex (Gender).

Kaplan-Meier Workflow - Step 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 - Step 2. Procedure – import data

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 - Descriptive analysis

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.
```

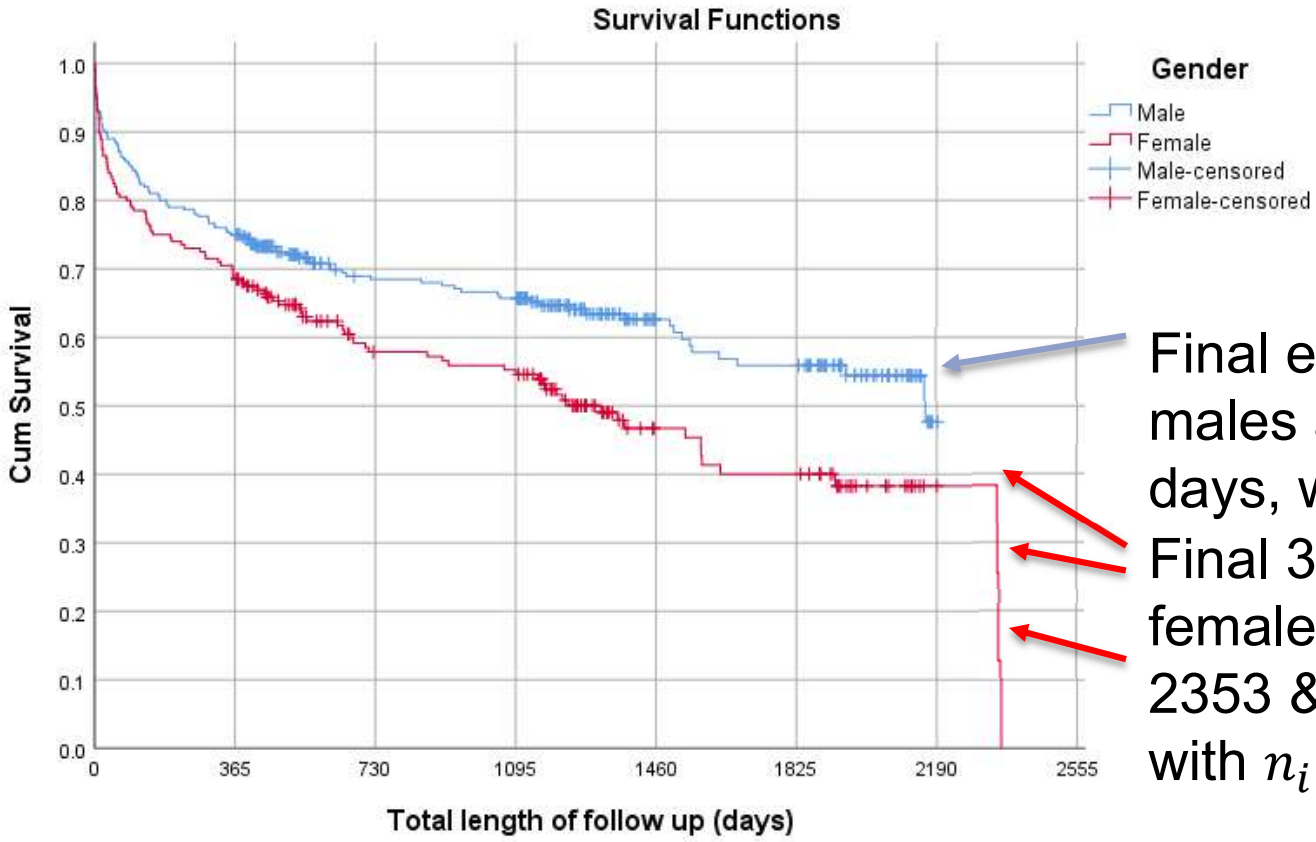
Have a look at the number of events in the dataset and the percentage censored.

Case Processing Summary

Gender	Total N	N of Events	Censored	
			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?

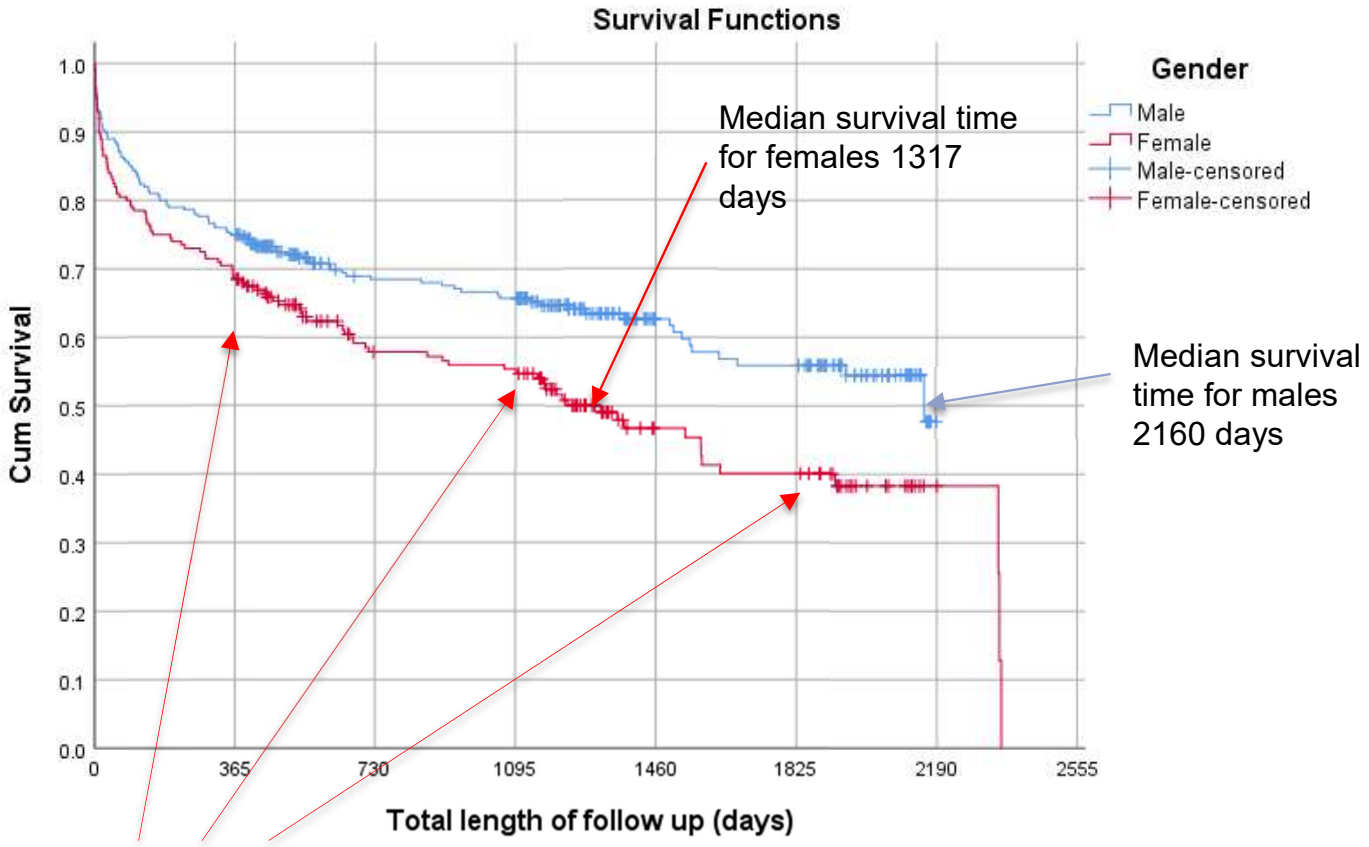
Kaplan-Meier 2. Procedure - Descriptive analysis



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 – Descriptive analysis



Kaplan-Meier 2. Procedure – Inferential analysis

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 - Step 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]

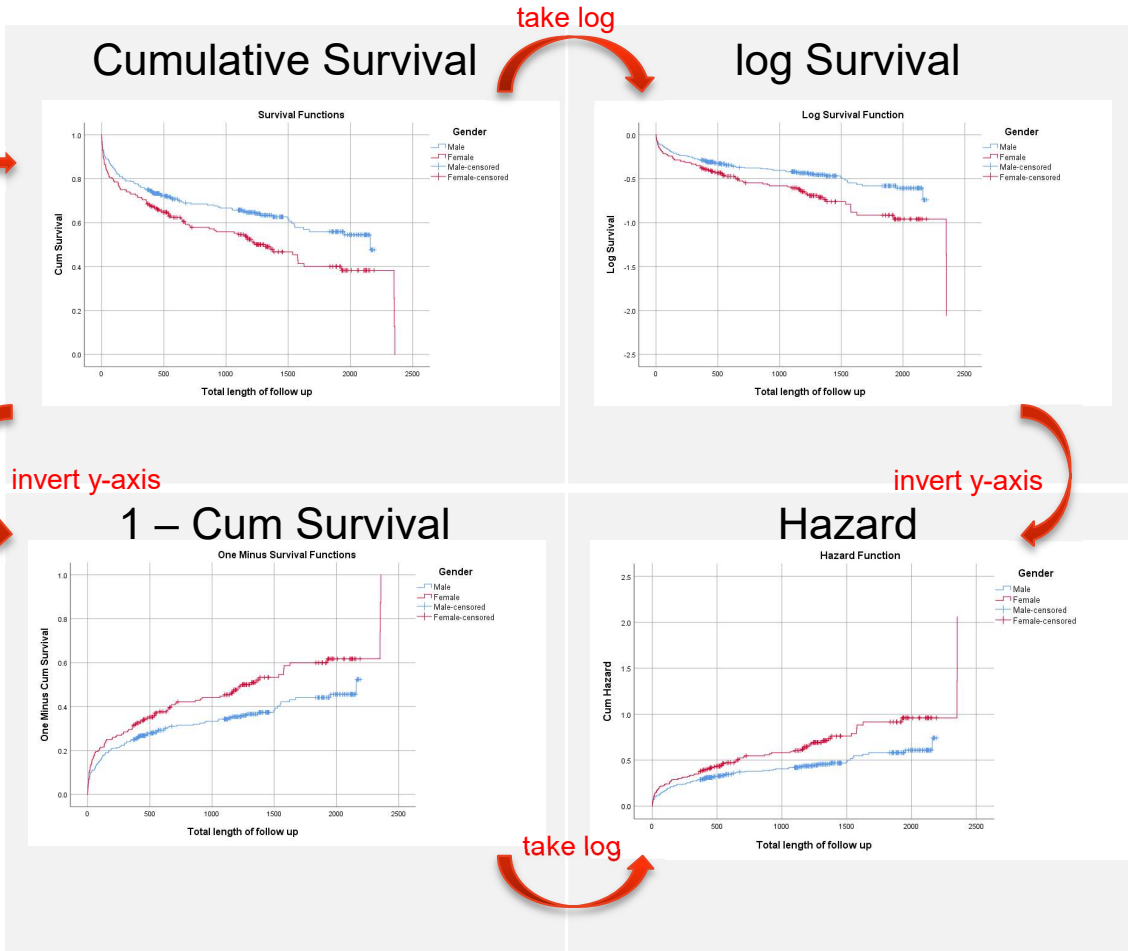
Median survival for females: 1317 days [95% CI 970-1664]

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

Kaplan-Meier – Step 3. Interpretation – 4 related plots

Include the cumulative Survival curve plot in your report.

The Survival function or Hazard function may also be represented by a variety of other plot options (as shown). Can depend on the nature of the event (positive/up or negative/down).

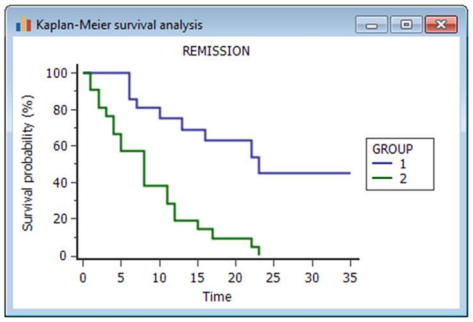


The log log survival curve can be used to assess the proportional hazards assumption: the two lines should be equidistant from each other over the entire time period.

T

Kaplan-Meier

Any questions?



Workflow: Steps in Survival Analysis

1. **Experimental and Analytical Design**
2. **Data cleaning**
3. **Exploratory Data Analysis** (EDA)
   - Survival data descriptive statistics
   - Kaplan-Meier procedure - plot
4. **Data Analysis aka inferential analysis**
   - Kaplan Meier associated tests (non-parametric)
   - Cox PH regression model (semi-parametric)
   - Advanced Models (parametric) and other model types
5. **Reporting for publication**

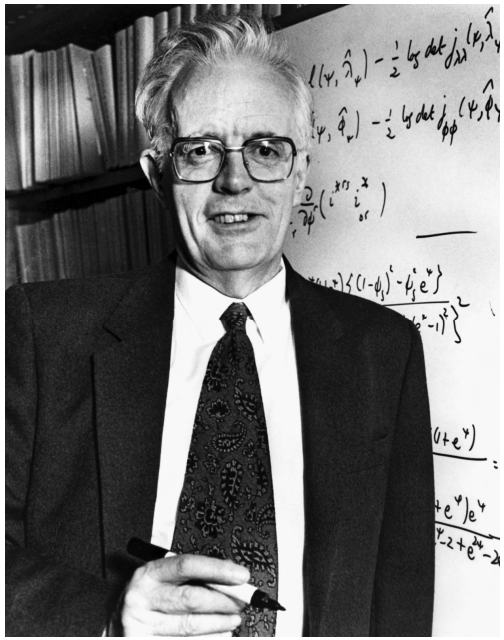
Cox Proportional Hazards Regression Introduction

- 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

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

(0.) Data cleaning + EDA - look at Kaplan-Meier survival curve

1. Fit a single Cox regression model
2. Check the model assumptions
3. Check goodness-of-fit
4. Interpret model parameters and reach a conclusion

Cox Proportional Hazards Regression Workflow

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

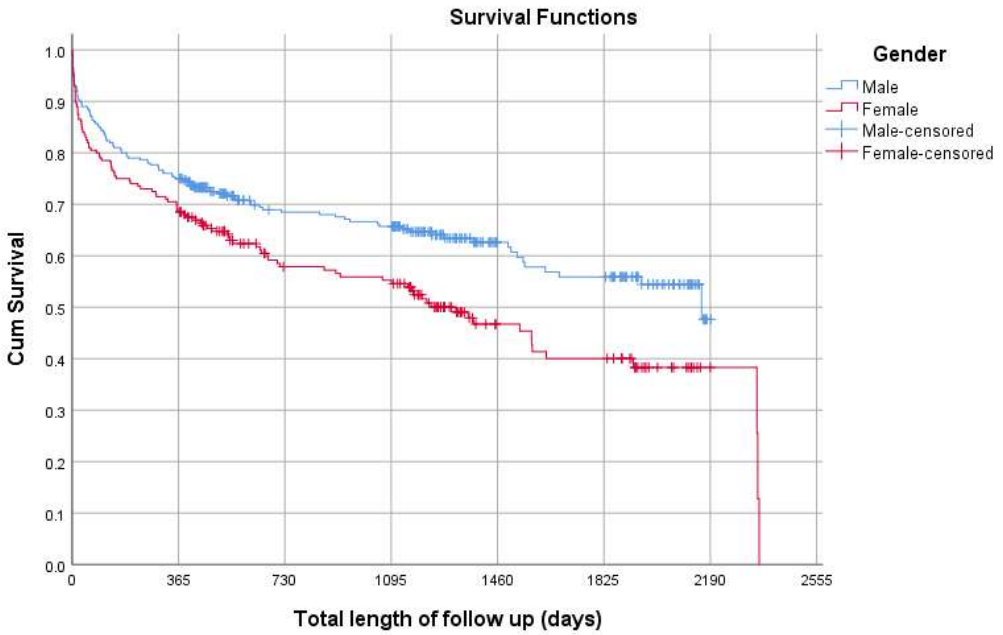
Example 1: Cox regression with one categorical predictor

What do we want to model? Let's start with a simple univariable model including Gender (Sex).

Cox PH Regression Workflow
Step 0: EDA- Kaplan-Meier

Step 0. EDA – Kaplan-Meier curve:

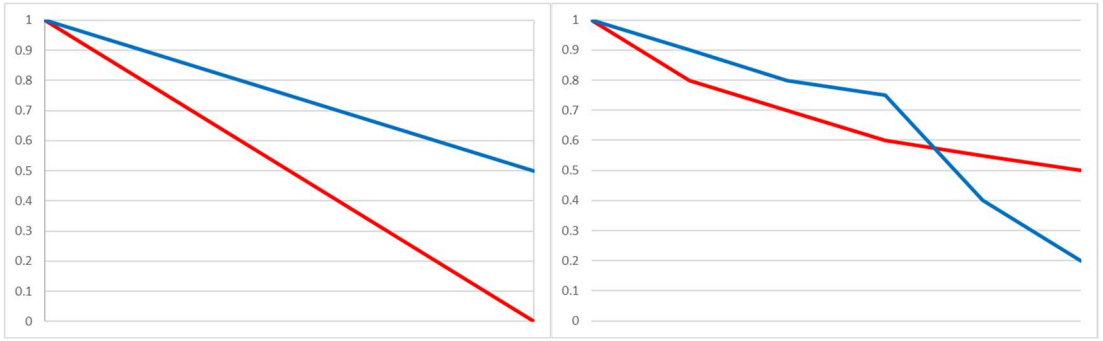
Have a look at the primary covariate of interest. Do the curves look proportional over the study period?



Cox Proportional Hazards Regression Workflow

What are we looking for in Step 0: EDA – Kaplan-Meier?

EDA – Kaplan-Meier survival curve
Proportional hazards

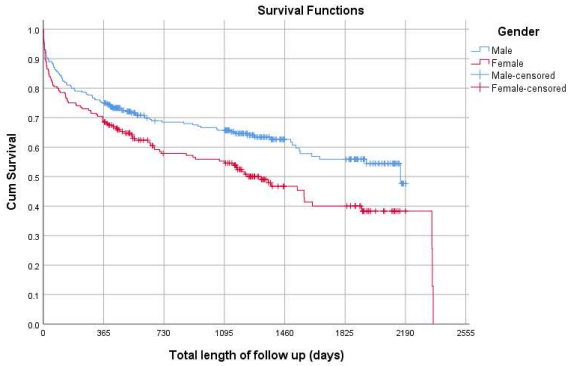


Constant rates for m & f
Unchanging Hazard Ratio over time

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

Step 1: Fit a Cox regression model with one predictor

Basic techniques are identical to those used in logistic regression

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

See Linear Models 2 and Statistical Model Building workshops for more information.

Cox Proportional Hazards Regression Workflow

1. Fit a Model
Run the Cox Regression procedure using your chosen software (SPSS shown)

Variables in the Equation

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

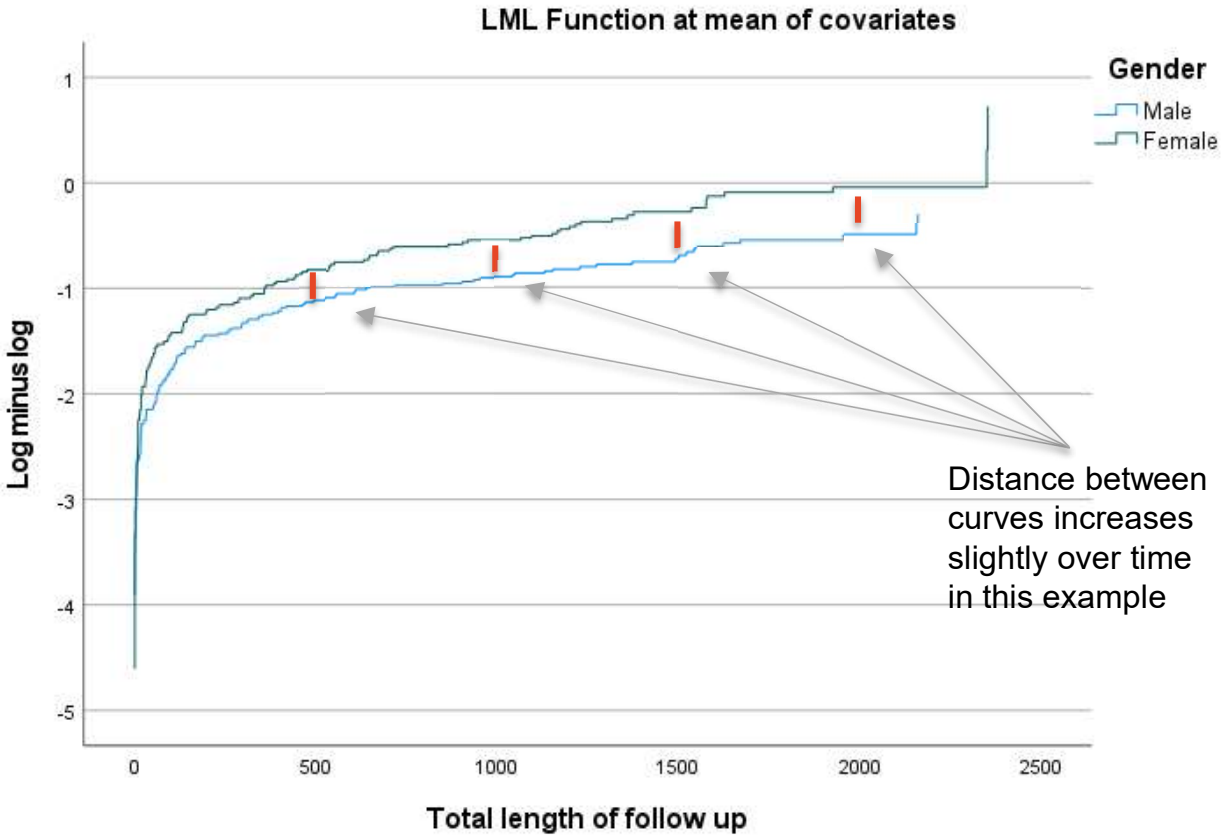
Gender is significant. We keep it in the model.
Hazard Ratio (Gender) = 1.464
(The odds of death occurring first for a female is ~1.5 compared to a male with a 95%)

Cox Proportional Hazards Regression Workflow

Step 2: Check the Model assumptions

Log minus Log curve can be used to assess
The proportional hazards assumption (that the hazard ratio stays the same over time). If the PH assumption is met, then the two curves will be equidistant apart along the length.

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.

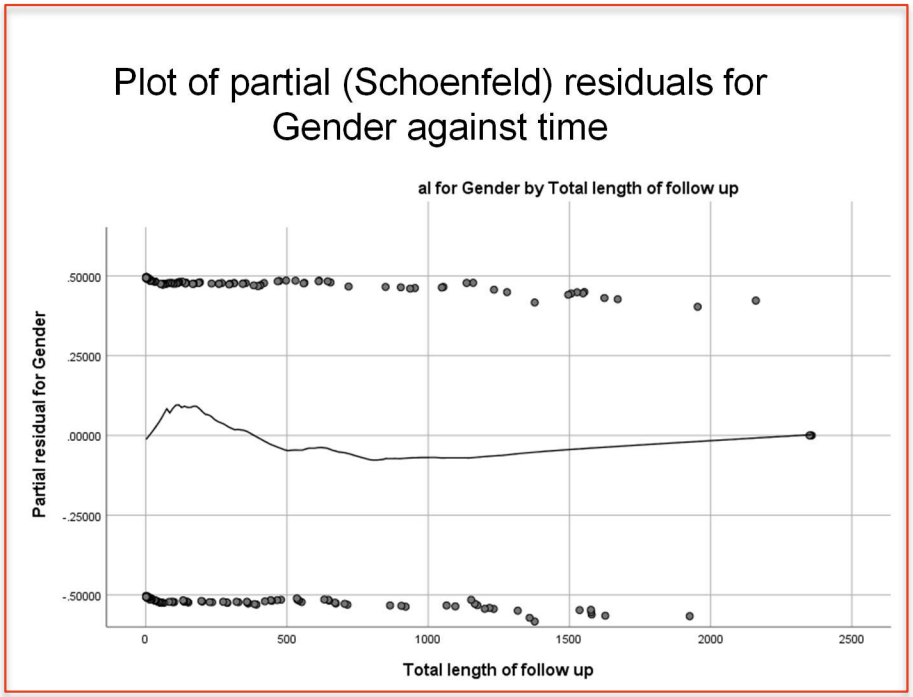


Cox Proportional Hazards Regression Workflow

2. 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.



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

Cox Proportional Hazards Regression Workflow

2. 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

Step 3: Check the Model Goodness-of-fit

Compared to the Null model with no predictors, the model with gender has an AIC of over 5 units lower, providing evidence that this model has a superior fit. See our Model Building Workshop for more information.

Variables in the model	-2 log L	AIC -2LogL+2q
Null	2455.2	2455.2
Gender	2447.6	2449.6

Goodness-of-Fit - other options:

- Compare observed and expected events (across G groups where G=integer (no. of events/40) – see 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

2. Check the model assumptions

Goodness of Fit

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

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

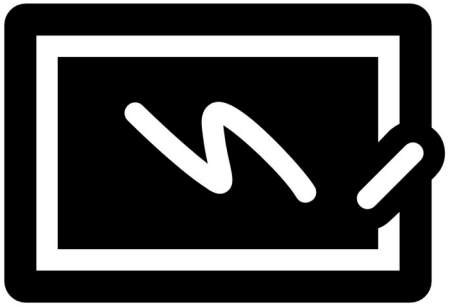
Where:

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

4. Interpret the model parameters and reach a conclusion

Variables in the Equation

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

Gender has a significant association with the hazard of dying (p=0.006). The Hazard Ratio for Gender = 1.464 and the reference category is male, so the odds of death occurring first for a female is ~1.5 compared to a male (95% confidence interval: 1.1-1.9).

Note: This is an observational study investigating associations. Randomised trials may have a stronger causal claim.

Example 2: Cox regression with a numeric predictor

Let’s model the numeric predictor BMI score.

Cox PH Regression Workflow
Step 0: EDA- Kaplan-Meier

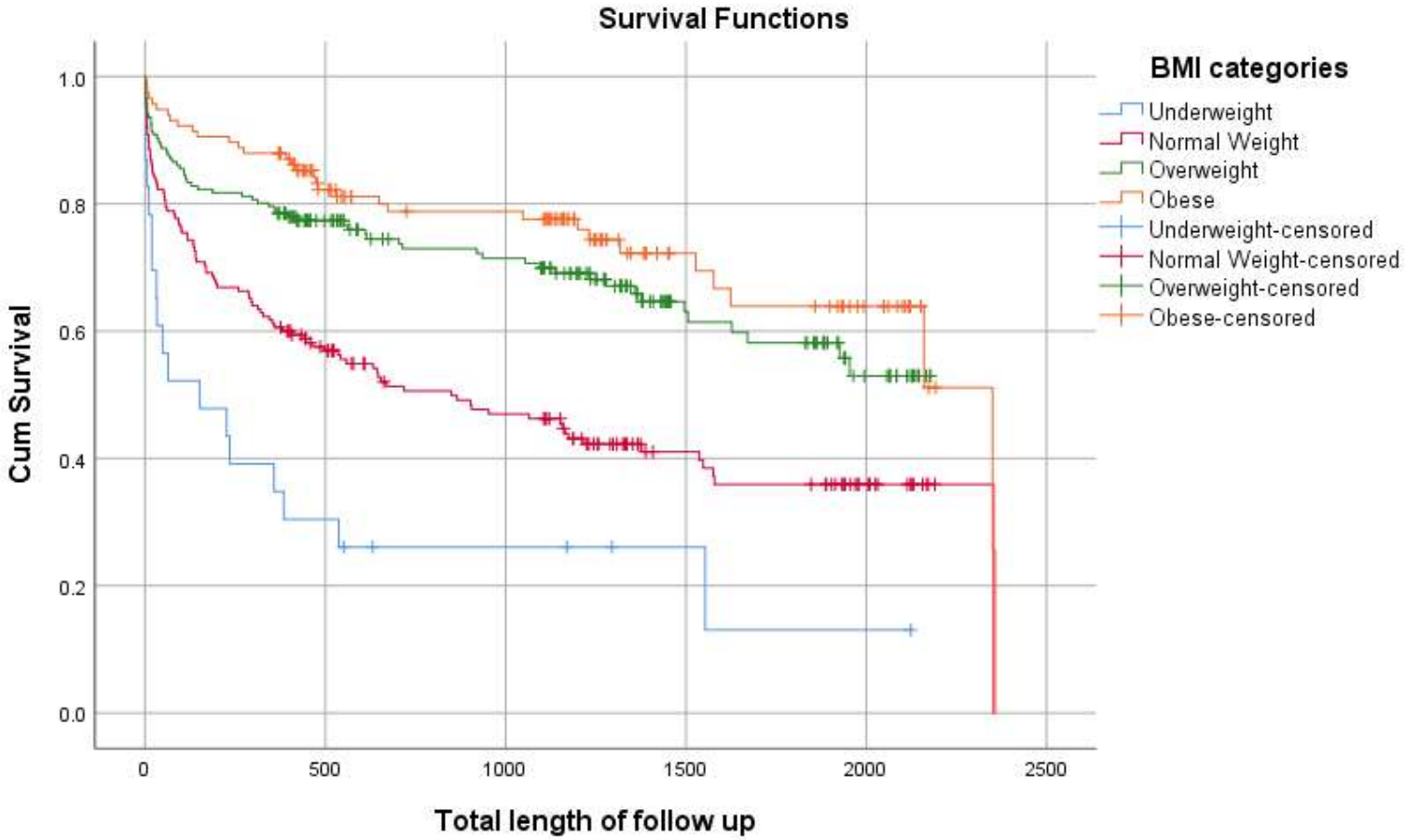
Step 0. EDA – EDA for a continuous, numeric predictor can be useful to **assess the assumption of linearity** (similar to Linear Models). To be able to do so for survival data we can categorise the numeric predictor and use Kaplan-Meier.

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

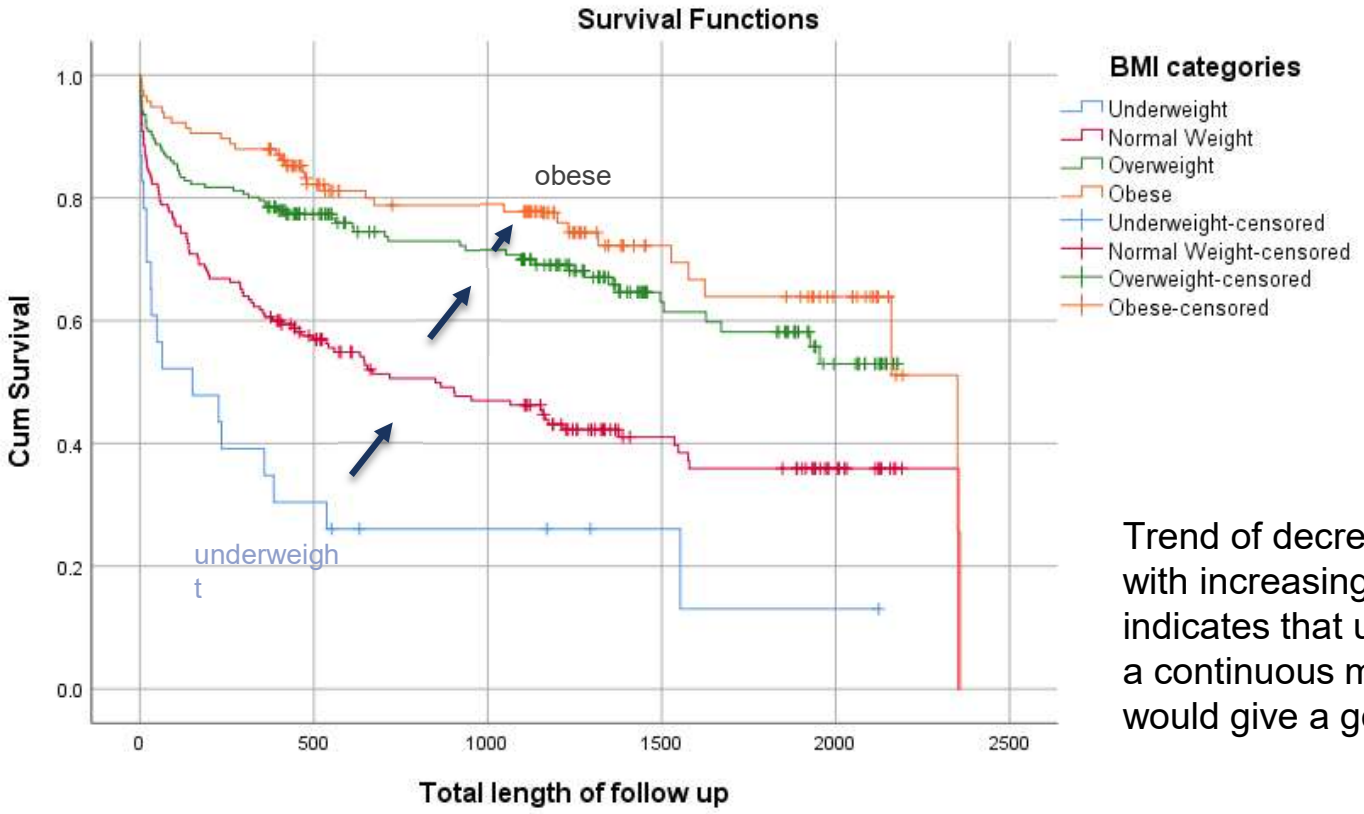
Question: Would you expect a linear relationship between BMI and survival?

Cox PH Regression Workflow - Step 0: EDA- Kaplan-Meier



Cox Proportional Hazards Regression Workflow

Step 0: EDA for numeric predictor + linearity assumption checking



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

1. Fit a Model
Run the Cox Regression procedure using your chosen software (SPSS shown)

		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
	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

The table shows HR estimates for BMI categories compared to “normal weight” as the reference category. Other choices of reference category may be preferred.
The BMI categorical variable is significant (p=0.007). Hazard Ratio (overweight compared to normal weight) = 0.614
We could run models with BMI as numeric and as categorical predictor and compare model fit and then make a decision which variable to use – the decision may also depend on ease of interpretation.

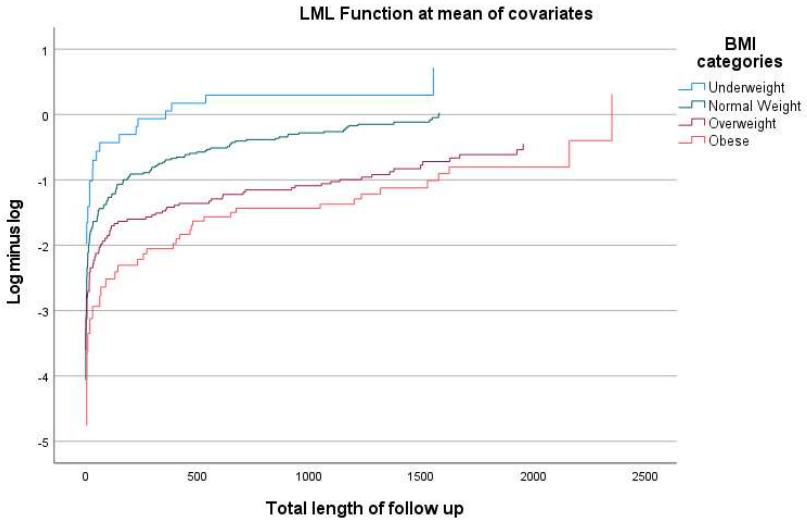
Cox Proportional Hazards Regression Workflow

Step 2: Check the Model assumptions

- Categorising the numeric predictor and checking the assumption of linearity graphically as shown under EDA – Kaplan Meier is one option to assess whether the proportional hazards assumption is met.
- 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

2. Check the model assumptions
Proportional hazards assumption: using Graphical methods
Look at the LML curves



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

Step 3: Check the Model Goodness-of-fit

Compared to the Null model with no predictors, the model with BMI has an AIC of over 56 units lower, providing evidence that this model has a superior fit (and BMI is a stronger predictor than Gender). See our Model Building Workshop for more information.

Variables in the model	-2 log L	AIC -2LogL+2q
Null	2455.2	2455.2
BMI	2407.0	2409.0

Cox Proportional Hazards Regression Workflow

4. Interpret the model parameters and reach a conclusion

		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
	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

BMI has a significant association with the hazard of dying (p=0.007). The odds of death occurring first for someone overweight are 1.6 less compared to a normal weight person (1/0.614= 1.629, the HR is protective). (95% confidence interval: 1.2-2.2).

Multivariable models with many predictors

Cox Proportional Hazards Regression Workflow

Other training and resources:
Attend our Statistical Model Building workshop for a more complete overview of this topic and for building multivariable models with many predictors.

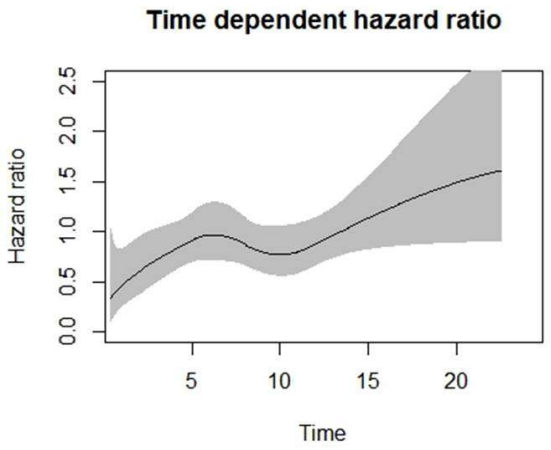
For building multivariable models (models that fit many predictor variables simultaneously, including interaction terms and confounders), similar considerations as for Linear Models apply to Survival Analysis.

Time dependent covariates in Cox PH regression:

Can use time-dependent covariates 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

Example plot:



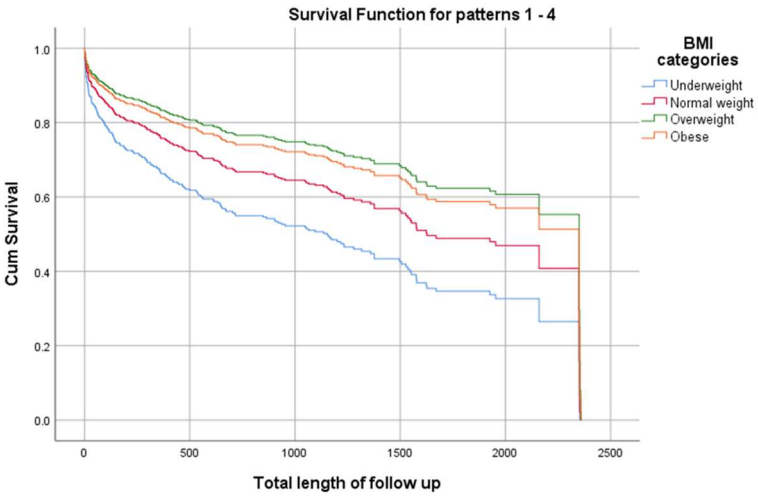
Add an interaction term into the Cox regression that include the "time" variable. This is **"Cox Regression with time-dependent covariates"**

- We will test the Time*BMI_cat interaction term in our model – SPSS defaults to call this term "T_Cov_": The Model output below shows that the interaction term is not significant so we can say that BMI_cat is time independent.

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
T_COV_	.000	.000	1.632	1	.201	1.000	1.000	1.000

Cox Proportional Hazards Regression

Any questions?



Cox, Box or Fox?

Workflow: Steps in Survival Analysis

1. **Experimental and Analytical Design**
2. **Data cleaning**
3. **Exploratory Data Analysis** (EDA)
   - Survival data descriptive statistics
   - Kaplan-Meier procedure - plot
4. **Data Analysis aka inferential analysis**
   - Kaplan Meier associated tests (non-parametric)
   - Cox PH regression model (semi-parametric)
   - Advanced Models (parametric) and other model types
5. **Reporting for publication**

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 (e.g. Exponential, Weibull, etc)- useful for prediction as these models make strong assumptions about the rate of survival over time

4. Frailty models – allows clustering to be modelled with a random effect (like in Mixed Models)

5. Competing Risks models – partitions event types

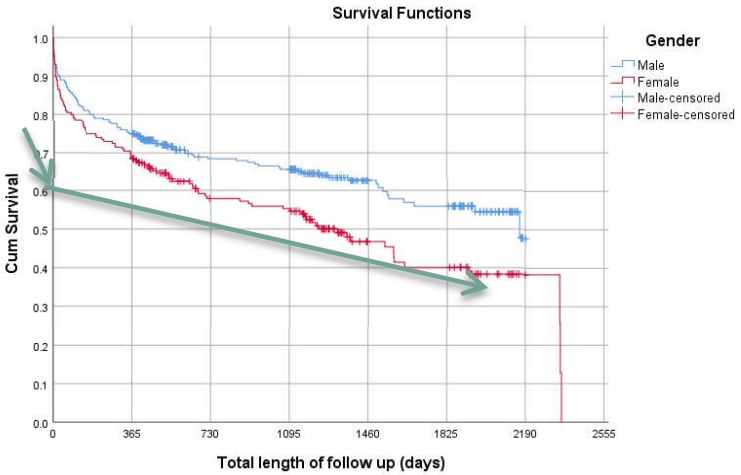
6. Discrete Time model using logistic regression – used when time is measured discretely with only a few values possible

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.

Workflow: Steps in Survival Analysis

1. **Experimental and Analytical Design**
2. **Data cleaning**
3. **Exploratory Data Analysis** (EDA)
   - Survival data descriptive statistics
   - Kaplan-Meier procedure - plot
4. **Data Analysis aka inferential analysis**
   - Kaplan Meier associated tests (non-parametric)
   - Cox PH regression model (semi-parametric)
   - Advanced Models (parametric) and other model types
5. **Reporting for publication**

Survival Analysis Workflow

5. Reporting for publication

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.

Survival Analysis Workflow

5. Interpret the model

Hazard Ratios – from WHAS example: how to report

- Categorical predictor:

The mortality hazard for females is 1.2 times [95% CI: 0.92-1.62] that of males.

- Numeric predictor:

The mortality hazard is increased by 6.3% [95% CI: 4.9-7.6%] for each additional year of age of the patient.

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

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

Useful Checklists for design and reporting in a publication

Reporting guidelines | EQUATOR Network, depends on study type, e.g.:

- CONSORT for Randomised trials
- STROBE for Observational studies

References - Software

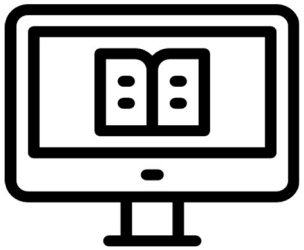
Software	accessibility	features
SPSS	Available to USyd staff and students	Kaplan Meier Cox Regression
STATA	via subscription	Kaplan Meier Cox Regression Many advanced model options
R packages: survival, survminer, survPen	free open source	K-M, Cox Regression Huge variety of options in these and other 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

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.

References



VIDEOS

Marinstats lectures https://youtu.be/vX3I36ptrTU

Z Statistics: Survival Analysis — Z Statistics – includes helpful videos, including for life tables and Excel calculators for parametric survival and hazard functions

WEBSITES

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

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 at The University of Sydney

SIH

- Statistical Resources website: containing our workshop slides and our favourite external resources (including links for learning R and SPSS).
- Hacky Hour: an informal monthly meetup for getting help with coding or using statistics software.
- 1on1 Consults can be requested on our website or here (click on the big red 'contact us' link).

SIH Workshops

- Create your own custom programs tailored to your research needs by attending more of our Statistical Consulting workshops. Look for the statistics workshops on our training page or on our Training calendar.
- Sign up to our mailing list to be notified of upcoming training.

Other

- Linkedin Learning

A reminder: Acknowledging SIH

- All University of Sydney resources are available to 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 for use of workshops and workflows:

- *"The authors acknowledge the Statistical workshops and workflows provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."*

We value your feedback

- We want to hear about you and whether this workshop has helped you in your research. What worked and what didn't work.
- We actively use the feedback to improve our workshops.
- Completing this survey really does help us and we would appreciate your help! It only takes a few minutes to complete (promise!)
- You will receive a link to the anonymous survey by email.

End of workshop

Kathrin Schemann
Senior Statistical Consultant

kathrin.schemann@sydney.edu.au

Statistical Consulting Unit
Sydney informatics Hub

sydney.edu.au/research/facilities/sydney-informatics-hub.html