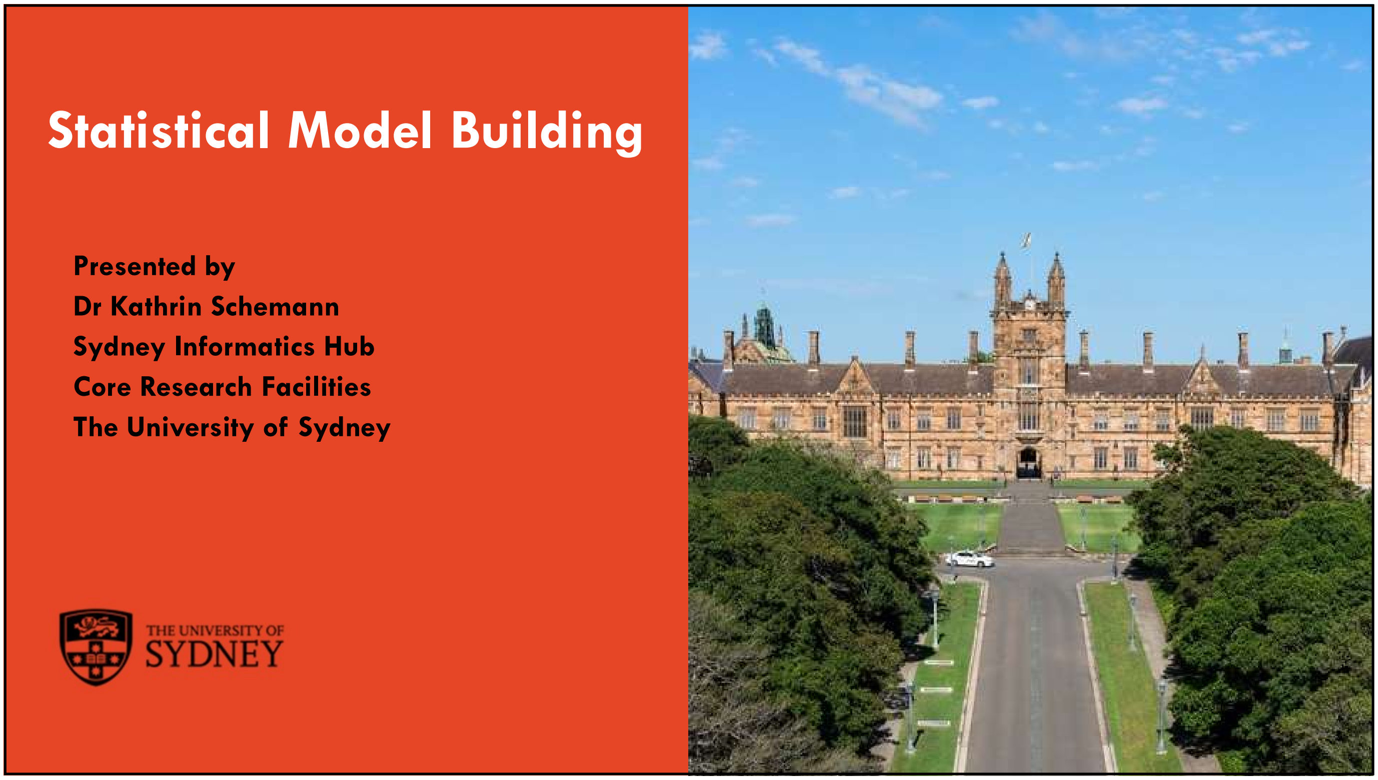
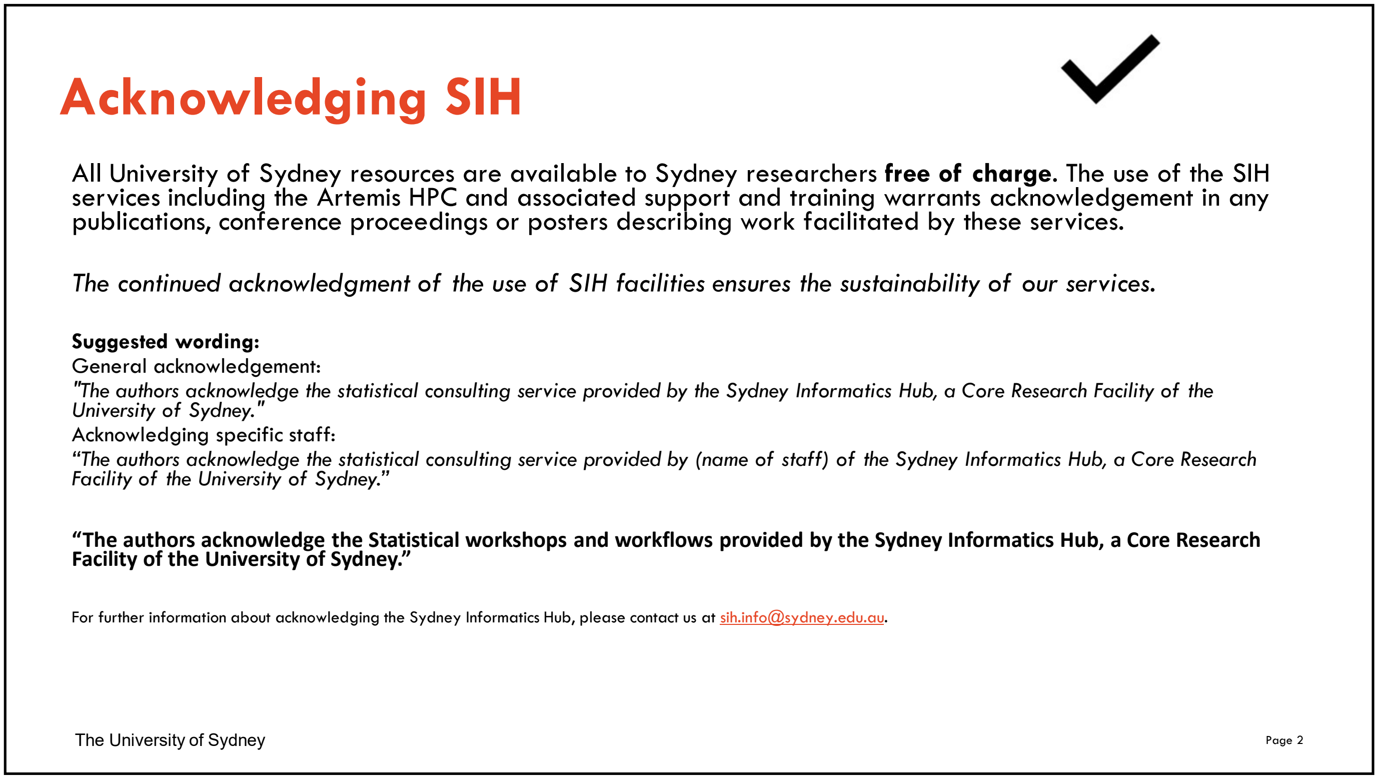


Statistical Model Building

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

1



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 statistical consulting service provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."
Acknowledging specific staff:
“The authors acknowledge the statistical consulting service provided by (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney.”

“The authors acknowledge the Statistical workshops and workflows provided by 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.

2

How to use this workshop

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 programmes tailored to their specific needs.*** This is achieved through:

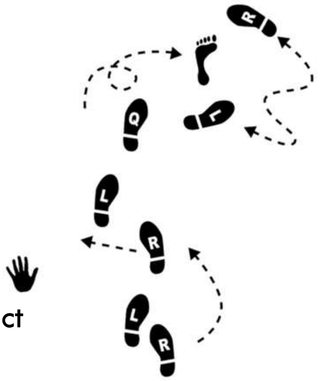
- **Short 90 minute workshops,** acknowledging researchers rarely have time for long multi day workshops.
- Providing **statistical workflows appliable in any software,** that give **practical step by step instructions which researchers return to when analysing and interpreting their data or designing 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 https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training.html#stats

3

Research Workflows

- **So… what is a workflow?**
  - The process of doing any statistical analysis follows the same general "shape".
  - We provide a general research workflow, and a specific workflow for each major step in your research (currently **experimental design, power calculation, analysis using linear models/survival/multivariate/survey methods)**
  - You will need to tweak them to your needs
- **Why do we need a research workflow?**
  - As researchers we are motivated to find answers *quickly*
  - But we need to be *systematic* in order to
    - Find the right method
    - Use it correctly
    - Interpret and report our results accurately
  - The payoff is huge, we can avoid mistakes that would affect the quality of our work *and* get to the answers sooner



4

Using this workshop after today

These slides should be used after the workshop as reference material and include these **workflows for you to follow**

- Todays workshop gives you the **statistical workflow**, which is software agnostic in that they can be applied in any software.

- There may also be accompanying **software workflows** that show you how to do it. We won't be going through these in detail. But if you have problems we have a monthly hacky hour where people can help you.

1on1 assistance You can request a consultation for more in-depth discussion of the material as it relates to your specific project. Consults can be requested via our Webpage (link is at the end of this presentation)

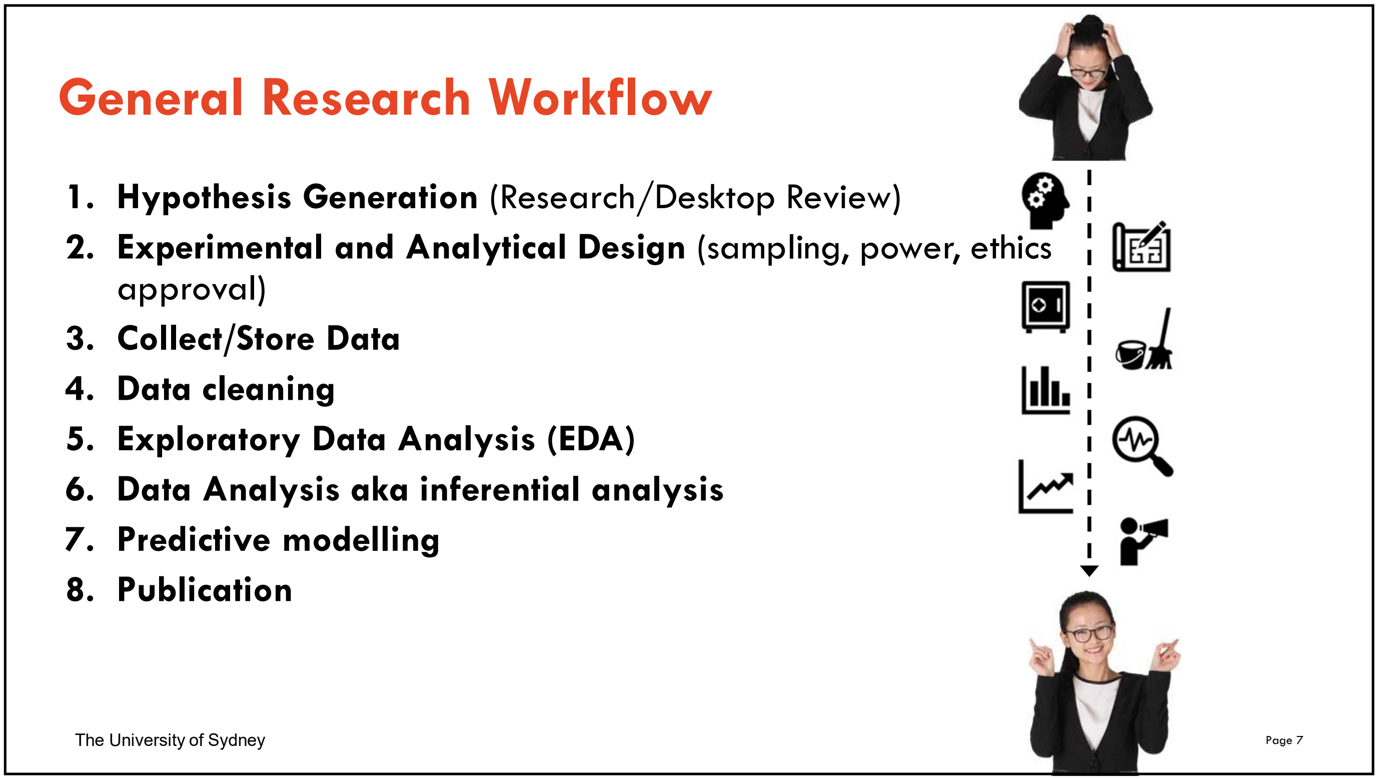
During the workshop

Ask short questions or clarifications during the workshop. 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 Question

- **A wild boar is coming towards you at 200mph. Do you:?**
  - A. Ask it directions
  - B. Wave a red flag
  - C. Wave a white flag
  - D. Begin preparing a trap



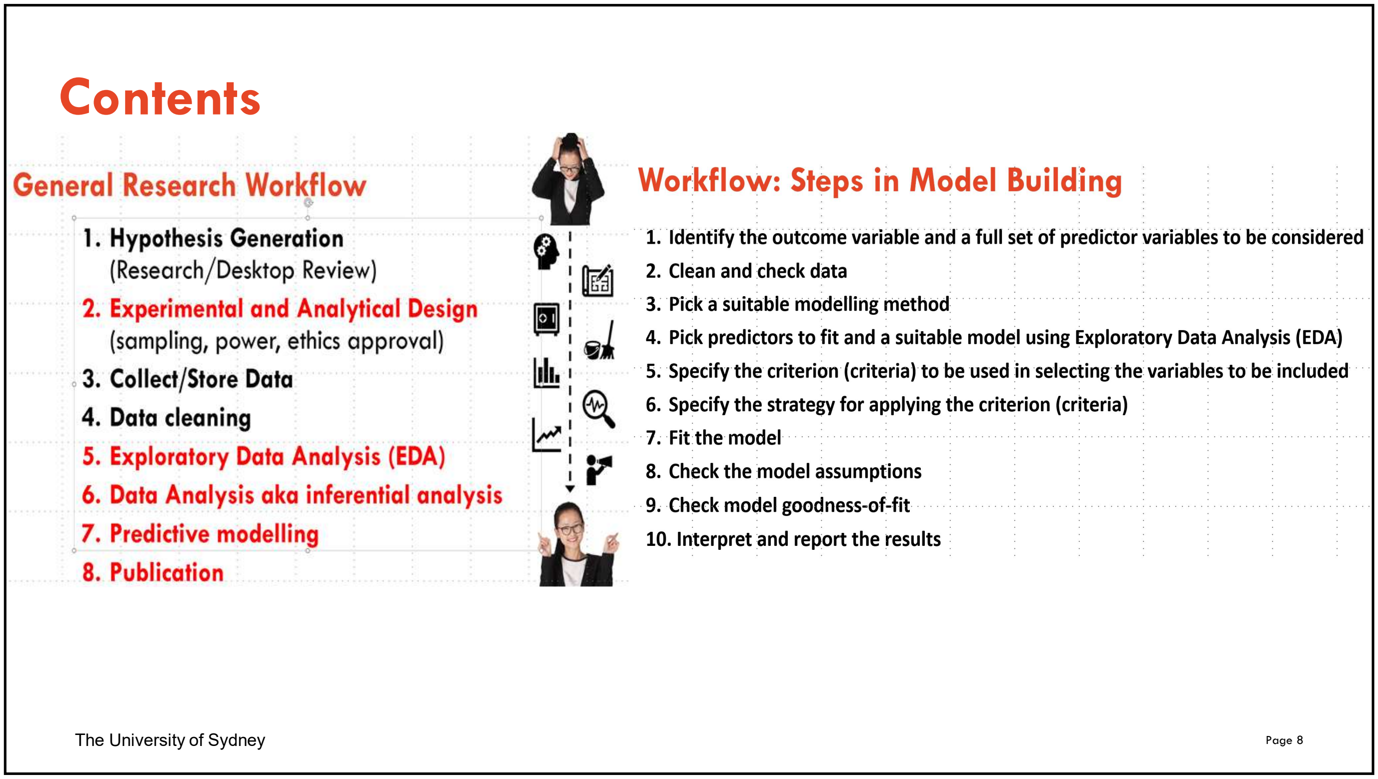
General Research Workflow

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

The University of Sydney

Page 7

7



Contents

General Research Workflow

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

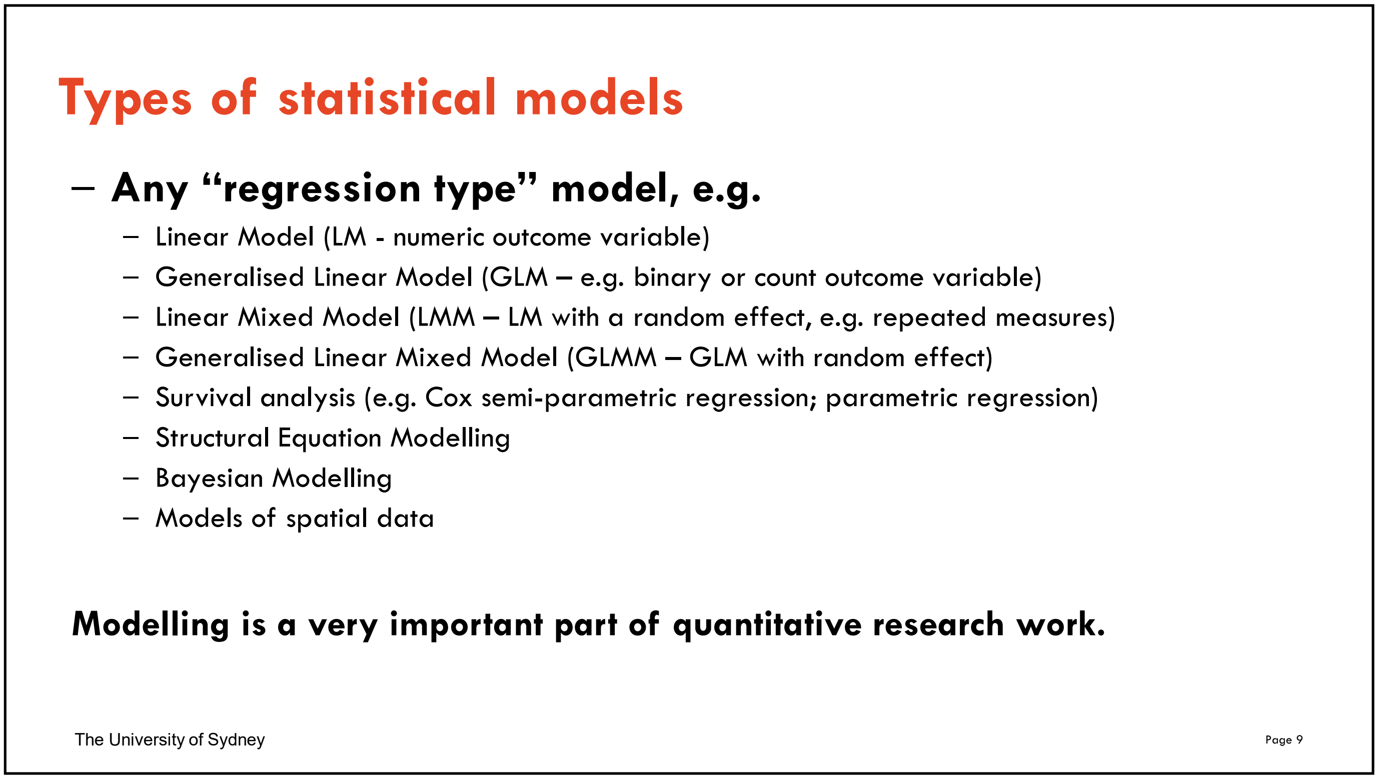
Workflow: Steps in Model Building

1. Identify the outcome variable and a full set of predictor variables to be considered
2. Clean and check data
3. Pick a suitable modelling method
4. Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)
5. Specify the criterion (criteria) to be used in selecting the variables to be included
6. Specify the strategy for applying the criterion (criteria)
7. Fit the model
8. Check the model assumptions
9. Check model goodness-of-fit
10. Interpret and report the results

The University of Sydney

Page 8

8

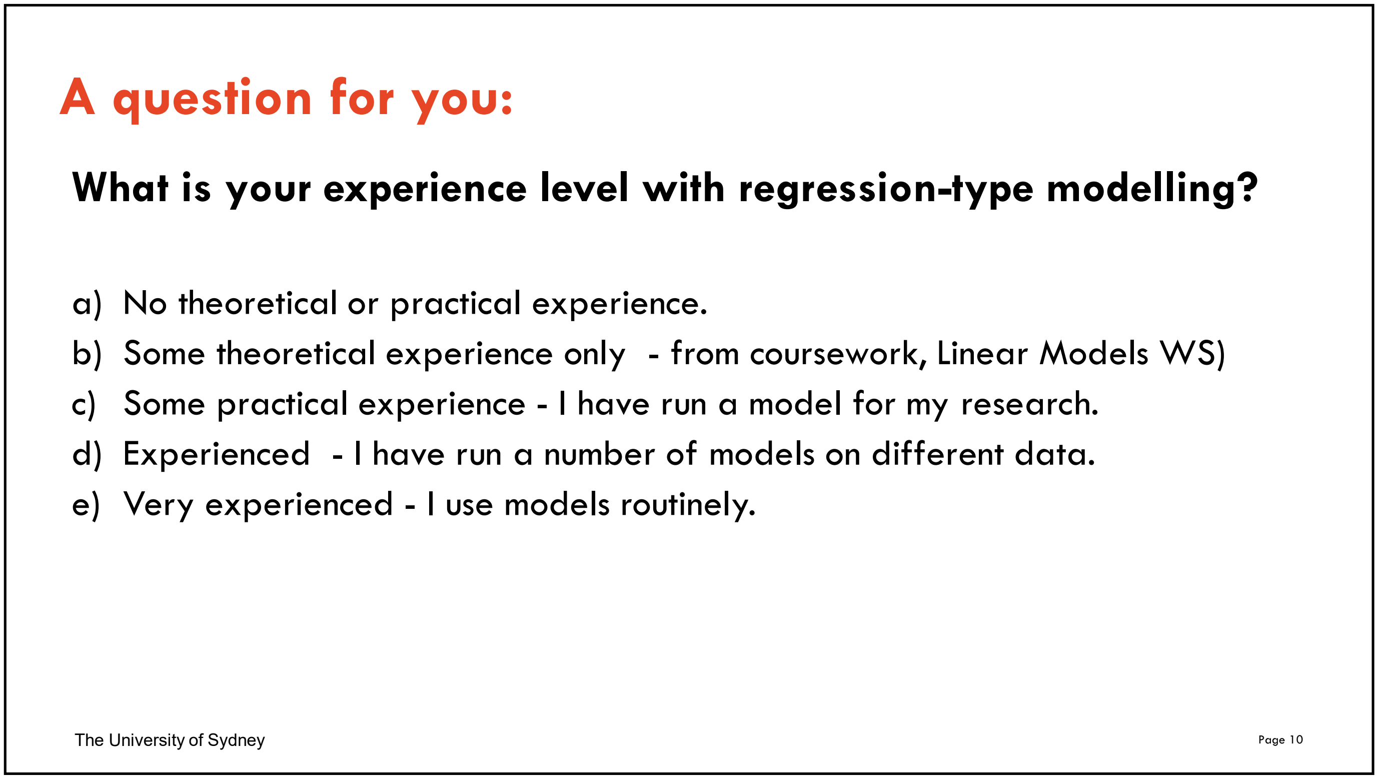


Types of statistical models

- Any "regression type" model, e.g.
  - Linear Model (LM - numeric outcome variable)
  - Generalised Linear Model (GLM – e.g. binary or count outcome variable)
  - Linear Mixed Model (LMM – LM with a random effect, e.g. repeated measures)
  - Generalised Linear Mixed Model (GLMM – GLM with random effect)
  - Survival analysis (e.g. Cox semi-parametric regression; parametric regression)
  - Structural Equation Modelling
  - Bayesian Modelling
  - Models of spatial data

Modelling is a very important part of quantitative research work.

9



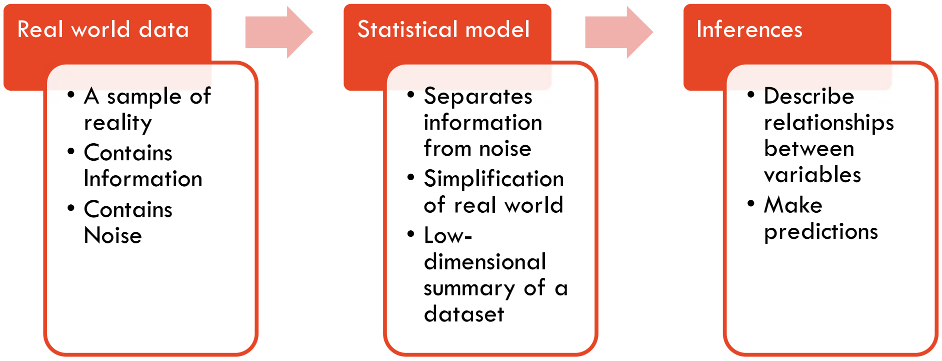
A question for you:

What is your experience level with regression-type modelling?

a) No theoretical or practical experience.
b) Some theoretical experience only - from coursework, Linear Models WS)
c) Some practical experience - I have run a model for my research.
d) Experienced - I have run a number of models on different data.
e) Very experienced - I use models routinely.

10

What is a model?



Real world data

- A sample of reality
- Contains Information
- Contains Noise

Statistical model

- Separates information from noise
- Simplification of real world
- Low-dimensional summary of a dataset

Inferences

- Describe relationships between variables
- Make predictions

"All models are wrong, but some are useful“ – George Box

Model Building – what is the ‘best fit’?

– **What is the purpose of a model?**

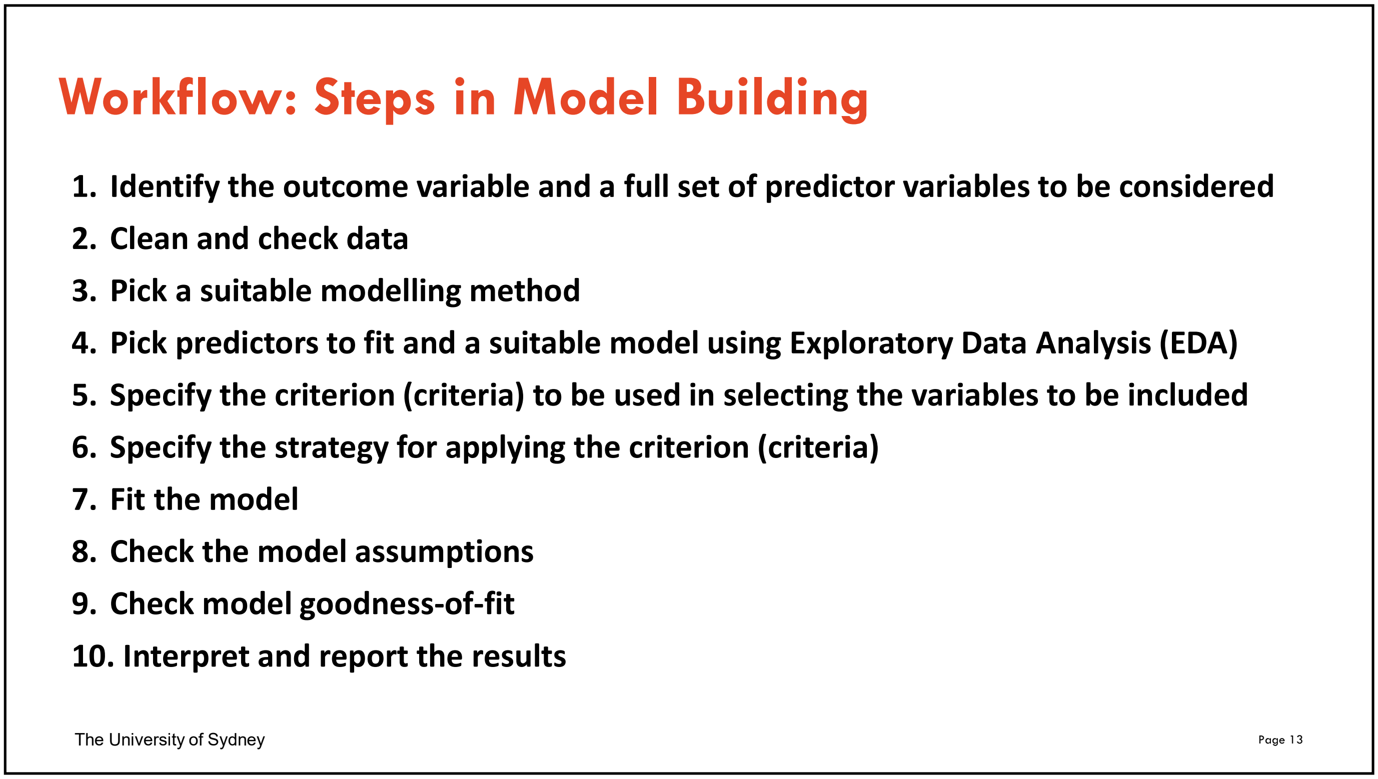
1. Hypothesis testing/ Inference – Interpret the relationships between predictors and the outcome
   - Gain knowledge - get the most precise estimates
   - Be careful of interaction and confounding
2. Prediction of future observations
   - Best model fit
   - Exclude predictors that are questionably related to outcome

– **Subject matter knowledge**

- Understand and consider the relationships between variables
- Consider ease of measurement and reliability of variables

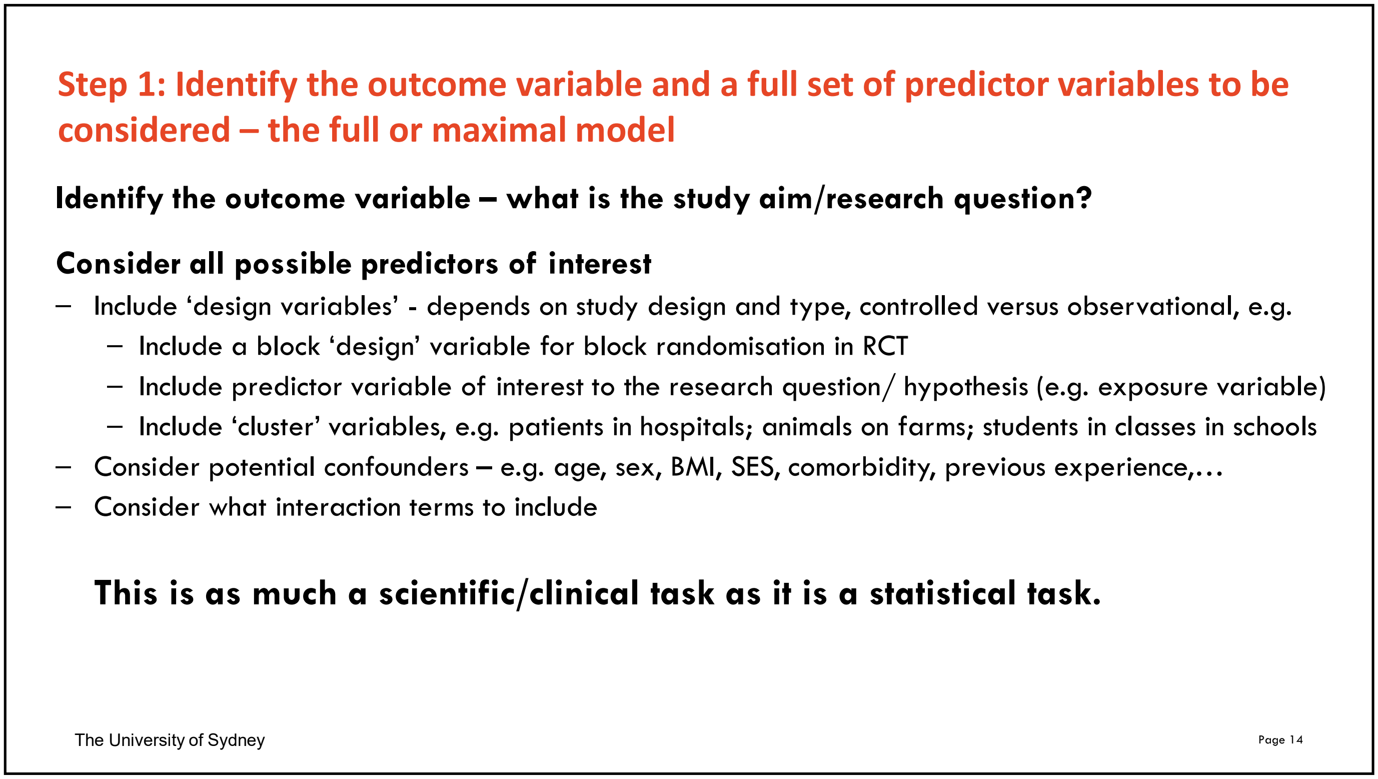
– **Parsimony (using as few predictors as required) versus fit**

- But include design variables/ known confounders



Workflow: Steps in Model Building

1. Identify the outcome variable and a full set of predictor variables to be considered
2. Clean and check data
3. Pick a suitable modelling method
4. Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)
5. Specify the criterion (criteria) to be used in selecting the variables to be included
6. Specify the strategy for applying the criterion (criteria)
7. Fit the model
8. Check the model assumptions
9. Check model goodness-of-fit
10. Interpret and report the results



Step 1: Identify the outcome variable and a full set of predictor variables to be considered – the full or maximal model

Identify the outcome variable – what is the study aim/research question?

Consider all possible predictors of interest

- Include ‘design variables’ - depends on study design and type, controlled versus observational, e.g.
  - Include a block ‘design’ variable for block randomisation in RCT
  - Include predictor variable of interest to the research question/ hypothesis (e.g. exposure variable)
  - Include ‘cluster’ variables, e.g. patients in hospitals; animals on farms; students in classes in schools
- Consider potential confounders – e.g. age, sex, BMI, SES, comorbidity, previous experience,…
- Consider what interaction terms to include

This is as much a scientific/clinical task as it is a statistical task.

Step 1: Identify the outcome variable and a full set of predictor variables to be considered – control of confounding

Potential confounders – e.g. age, sex, BMI, SES, comorbidity, previous experience, etc.

Control at study design stage:

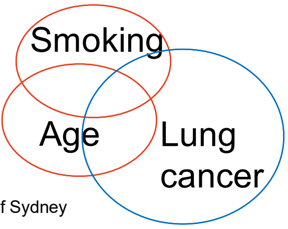
– Use exclusion (restricted sampling; e.g. only people without comorbidity)

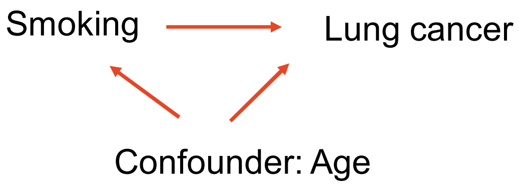
– Use stratification (divide sample into different groups)

– Use matching e.g. matched case-control study (matched on same comorbidity); blocking in RCT

At analysis stage:

– Use multivariable regression analysis for analytical control/adjustment




Smoking

Lung cancer

Confounder: Age

Step 1: Identify a full set of predictor variables - Interactions

Decide what interactions between predictors are to be considered

- Avoid 3-way interactions unless indicated by subject matter (difficult to interpret)
- Consider multiple testing adjustment if using large number of interactions – see **Linear Models 3 WS**
- Evaluate interactions graphically where possible.
- Rule of thumb: Having more than two interaction terms usually over-complicates the model and may not represent reality

Five general strategies for creating and evaluating interactions*:

1. Create and evaluate all 2-way interaction terms (ok for number of predictors <=8)
2. Create 2-way interactions among all predictors that are significant in the final main effects model (after model building)
3. Create 2-way interactions among all predictors found to have an unconditional (univariate) association with the outcome.
4. Create 2-way interactions only among pairs of variables which you suspect might interact (based on literature, expert knowledge etc), e.g. those involving the primary predictor of interest and important confounders.
5. Only create 2-way interaction terms that involve the predictor of interest.

*** May be adjusted for 'biologically plausible' interactions only.**

Step 1: Identify a full set of predictor variables - pitfalls

Over-fitting the data - too many predictors, not enough data/sample size.

Rule of thumb: one should have at least 10 data points for each estimate

– An intercept needs an estimate

– Each numeric or binary predictor needs an estimate

– Each dummy variable needs an estimate, i.e. a categorical variable with 4 categories needs 3 estimates (one category will be the reference group)

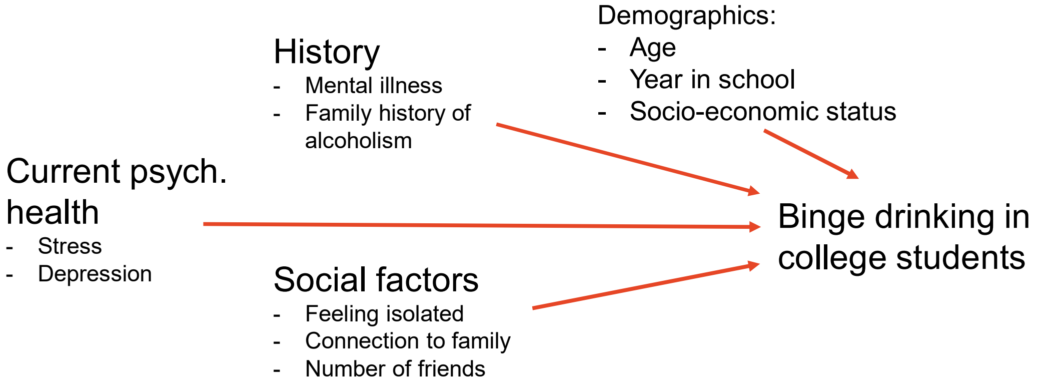
– Interaction terms need estimates – if an interaction term is included in the model the main effects need to also be included in order to be able to interpret the interaction estimates

If the goal is to create knowledge then we need our estimates to be as precise as possible – too little data will lead to large confidence intervals.

A large maximum model should include all potentially important predictors but it increases the chance of multi-collinearity, unstable estimates and finding spurious associations that are not important in the real world or are difficult to interpret. Consider a focused study design collecting high quality data on far fewer predictors versus administrative 'big' data not collected for research purposes.

Step 1: Identify a full set of predictor variables – other strategies

- Consider the causal structure of your data
- Consider drawing a causal diagram (Directed acyclic graph – DAG, see appendix)
- Group variables into logical clusters (for large numbers of predictor variables you may consider working in clusters for model building)



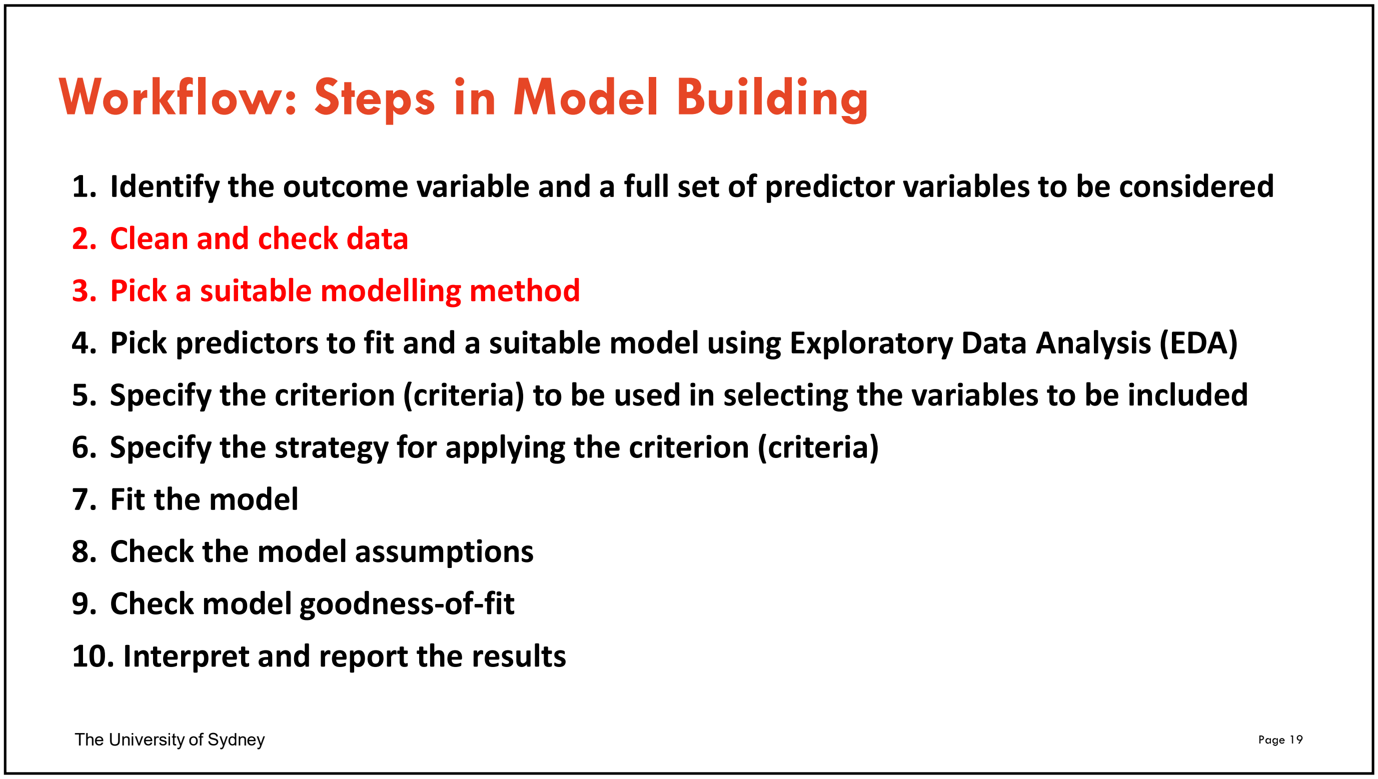
History
- Mental illness
- Family history of alcoholism

Demographics:
- Age
- Year in school
- Socio-economic status

Current psych. health
- Stress
- Depression

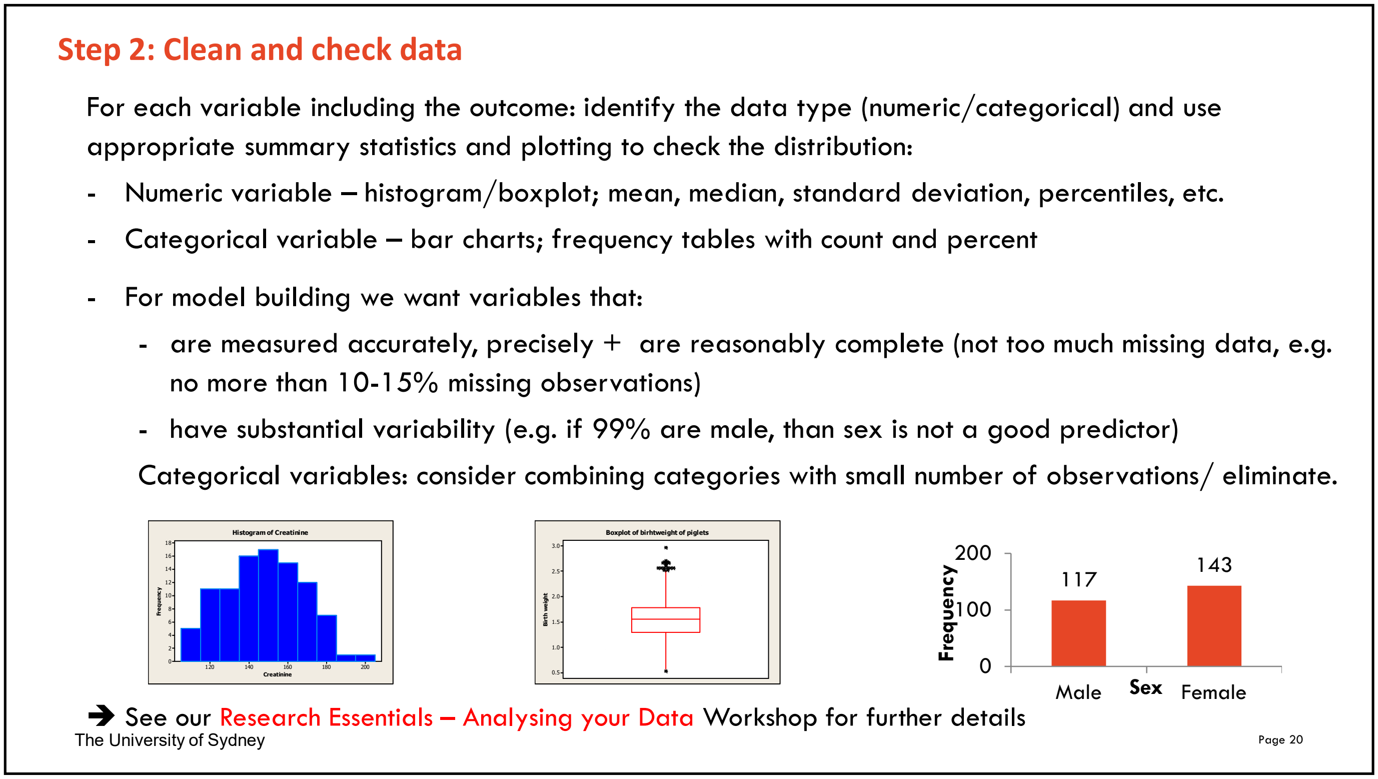
Binge drinking in college students

Social factors
- Feeling isolated
- Connection to family
- Number of friends



Workflow: Steps in Model Building

1. **Identify the outcome variable and a full set of predictor variables to be considered**
2. **Clean and check data**
3. **Pick a suitable modelling method**
4. **Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)**
5. **Specify the criterion (criteria) to be used in selecting the variables to be included**
6. **Specify the strategy for applying the criterion (criteria)**
7. **Fit the model**
8. **Check the model assumptions**
9. **Check model goodness-of-fit**
10. **Interpret and report the results**



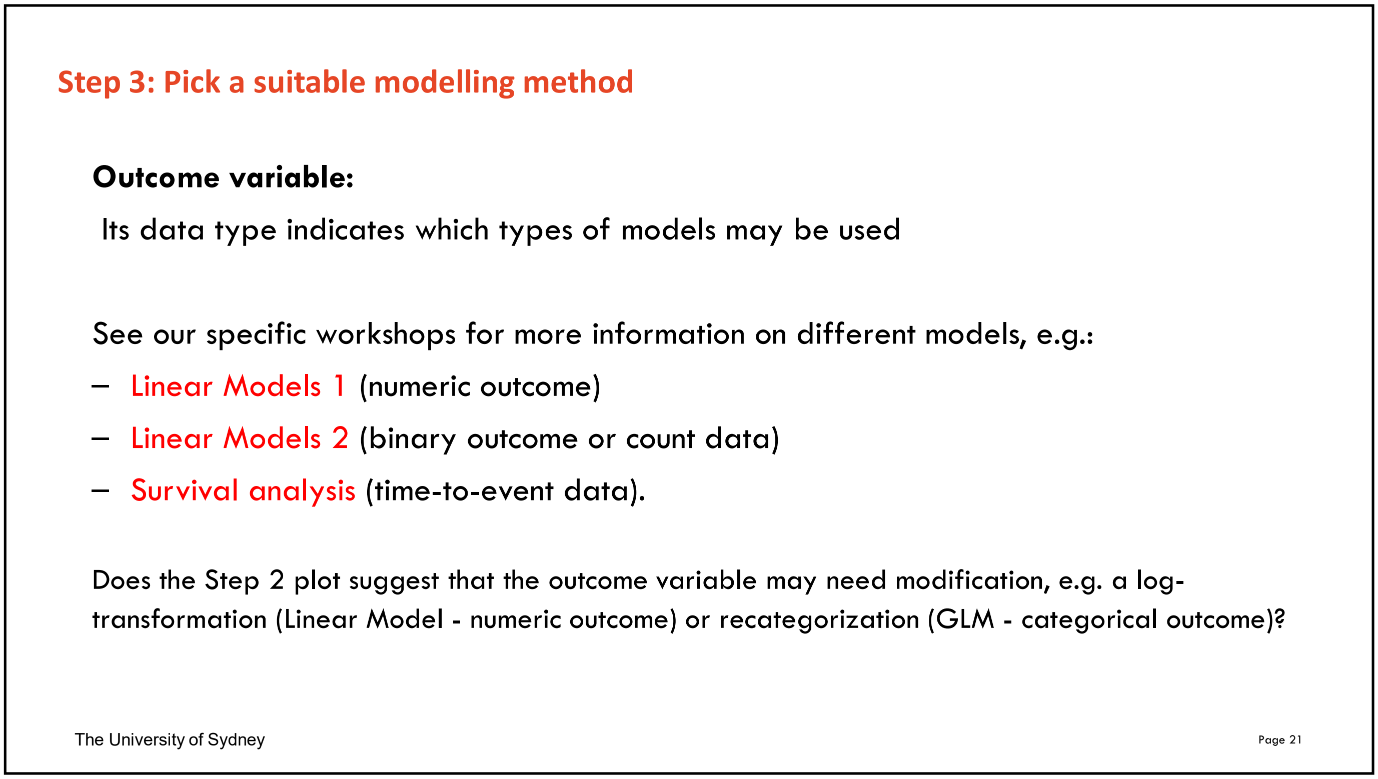
Step 2: Clean and check data

For each variable including the outcome: 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
- For model building we want variables that:
  - are measured accurately, precisely + are reasonably complete (not too much missing data, e.g. no more than 10-15% missing observations)
  - have substantial variability (e.g. if 99% are male, than sex is not a good predictor)

  Categorical variables: consider combining categories with small number of observations/ eliminate.

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



Step 3: Pick a suitable modelling method

Outcome variable:

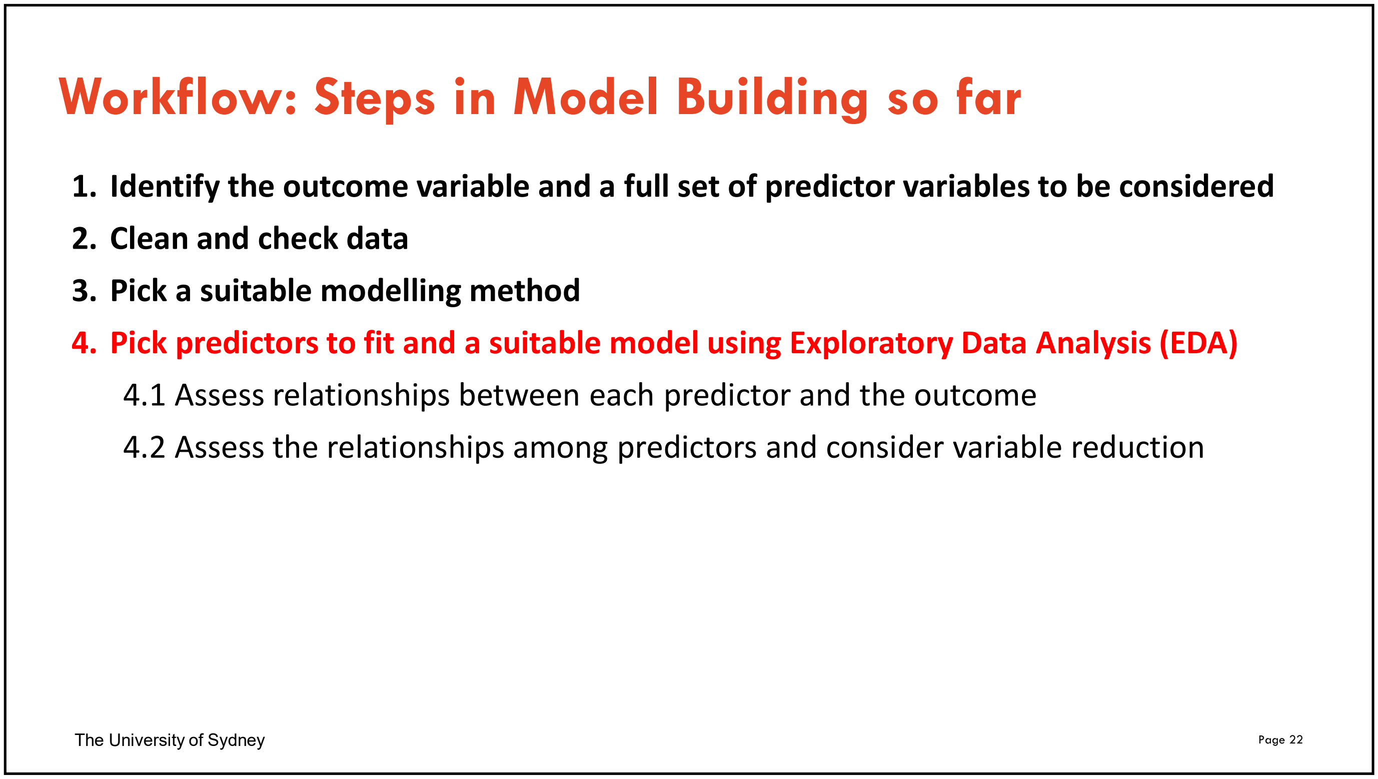
Its data type indicates which types of models may be used

See our specific workshops for more information on different models, e.g.:

- Linear Models 1 (numeric outcome)
- Linear Models 2 (binary outcome or count data)
- Survival analysis (time-to-event data).

Does the Step 2 plot suggest that the outcome variable may need modification, e.g. a log-transformation (Linear Model - numeric outcome) or recategorization (GLM - categorical outcome)?

21



Workflow: Steps in Model Building so far

1. **Identify the outcome variable and a full set of predictor variables to be considered**
2. **Clean and check data**
3. **Pick a suitable modelling method**
4. **Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)**

    4.1 Assess relationships between each predictor and the outcome

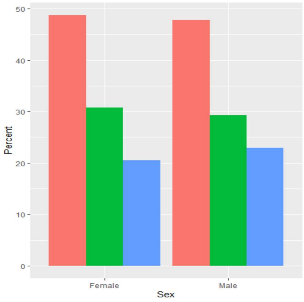
    4.2 Assess the relationships among predictors and consider variable reduction

22

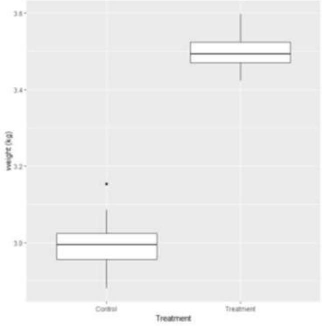
Step 4: Pick predictors to fit using Exploratory Data Analysis (EDA) 1

4.1 Assess relationships between each predictor and the outcome to select, modify + understand predictors

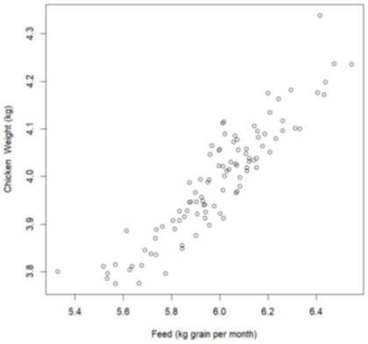
- plot the relationship of each predictor with the outcome



2 categorical variables
- side-by-side bar charts



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



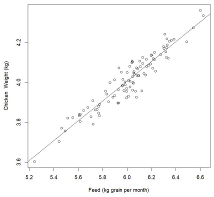
2 numeric variables
– xy scatter plot

Step 4: Pick predictors to fit using Exploratory Data Analysis (EDA)

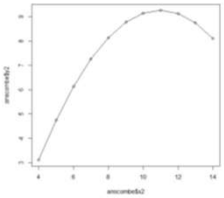
4.1 Assess relationships between each predictor and the outcome to select, modify + understand predictors

For numeric variables in linear regression – assess the assumption of linearity – is the relationship between outcome and predictor a line, a curve or something else?

For practical reasons we assess the model assumption of linearity before model building.



Linear Model (a straight line):
$y = b_0 + b_1x_1 + error$



Quadratic model (a curve):
$y = b_0 + b_1x_1 + b_2x_1^2 + error$

Alternatively, avoid the assumption by categorising the numeric predictor, but this loses information and may introduce bias. However, sometimes we are interested in specific categories for interpretation, e.g. BMI – underweight/normal weight/ overweight/obese.

Non-linear regression models

Add power terms to linear model with predictor x , e.g. x^2 or x^3 – this allows the regression line to follow a curve

- Complexity/ number of bends depends on the number of power terms
- Global influence – influenced by the whole dataset
  - May perform better with future data
  - Less sensitive to local disturbance and may miss features
  - May be heavily influenced by values at the extreme ends
  - Do not predict outside of data range
- avoid collinearity among new variables - centre variables by subtracting the mean (mean of new variable = 0) to create orthogonal (uncorrelated) polynomials → limited to positive integer power terms

Functional form of continuous predictor's 'best fit' can be determined by statistical significance, but for strong subject matter reasons may choose polynomial anyway, particularly for small dataset and confounding predictor

Fractional polynomials and piecewise functions/splines - even more flexibility in fit (and complexity!)

Workflow: Steps in Model Building so far

1. **Identify the outcome variable and a full set of predictor variables to be considered**
2. **Clean and check data**
3. **Pick a suitable modelling method**
4. **Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)**

   4.1 Assess relationships between each predictor and the outcome

   4.2 Assess the relationships among predictors and consider variable reduction

WHY is EDA important??

"Knowledge without practice is useless; practice without knowledge is dangerous."

Confucius

Consider results from a multivariable model for Math test score…

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.105	1.104	1.9	0.0569 .
IQ	0.997	0.006	151.9	<2e-16 ***
verbal.IQ	-9.995	0.067	-148.3	<2e-16 ***

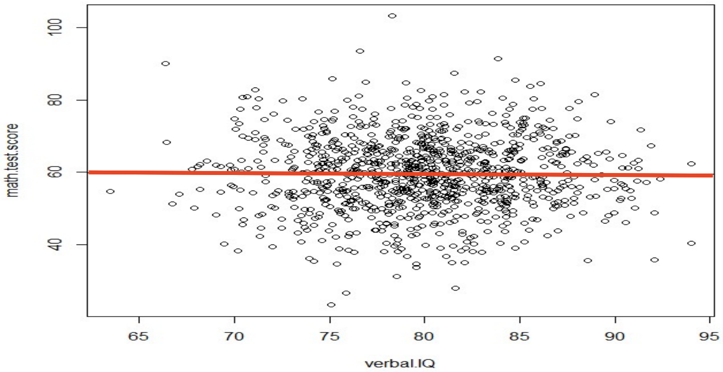
Multiple R-squared: 96%

$$y = b_0 + b_1 x_1 + b_2 x_2$$

Math test score = 2.1 + 1*IQ – 10*verbal.IQ

What is your interpretation? Is this a good model?

EDA for verbal IQ and Math test score:



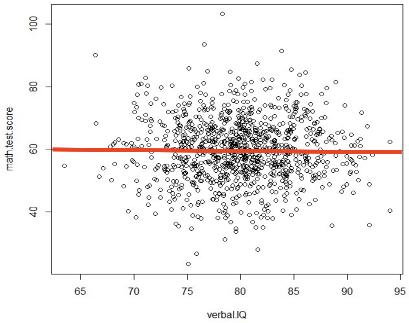
- **EDA:** correlation r=0 → there is no relationship!
- **MV Model:** on average with a 1 point increase in verbal IQ, Math test score decreases by 10 points!

Start simple, get complex: use Exploratory Data Analysis (EDA) to assess the relationships among predictors and consider variable reduction

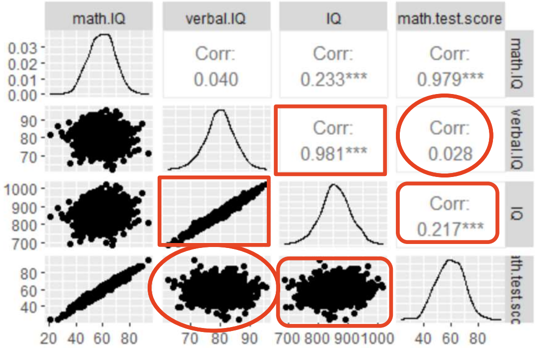
Numeric predictors: use xy scatter plot and conduct **pairwise correlation analysis**

'High' correlation is domain-specific, generally r>|0.7|,|0.8|or |0.9| but r<0.7 can be problematic too

Only use one of a pair of highly correlated variables for multivariable modelling – **WHY?**



Correlation coefficient r = 0



Correlation matrix

Simulated data

math.IQ <- rnorm(1000, 60, 10)
math.test.score <- math.IQ + rnorm(length(math.IQ),0,2)

verbal.IQ <- rnorm(1000, 80, 5)

IQ <- 10*verbal.IQ + math.IQ

- **math.test.score is math.IQ with a bit of noise**
- **IQ is a sum of verbal IQ and math.IQ (IQ:verbal.IQ r = 0.98)**
- **A model with verbal.IQ and IQ can re-arrange itself to:**

  Math.IQ = IQ – verbal.IQ

Example summary

Results:

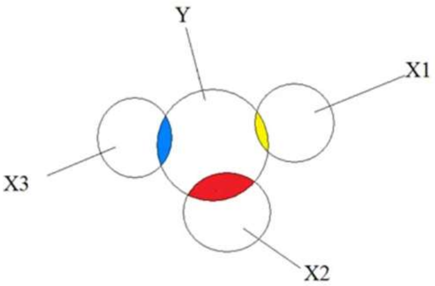
- *Verbal IQ* is not negatively correlated with *Math Test Score* as shown by the simple marginal model and the simulated data
- Yet a model with *Verbal IQ* and *IQ* suggests *Verbal IQ* is negatively corelated with *Math Test Score.*
- WHY: because *IQ=verbal IQ + math IQ meaning* a model which has both *IQ* and *verbal IQ* (which are correlated) can rearrange itself as *math IQ = IQ – verbal IQ.*
- This a reversal of *verbal IQ - verbal IQ* is **SUPRESSING** *IQ* (the part of IQ that is uncorrelated with Math results).

➔This is an example of the **reversal paradox** due to multi-collinearity – this may be called *Simpson's paradox*, *Lord's paradox*, and *suppression* - depending on whether the outcome and explanatory variables are categorical, continuous or a combination of both

➔This is why we always look for multi-collinearity during the Exploratory Data Analysis using a scatterplot matrix to learn about the relationships between variables.

➔**Start simple, get complex!**

Relationships between variables - Correlation

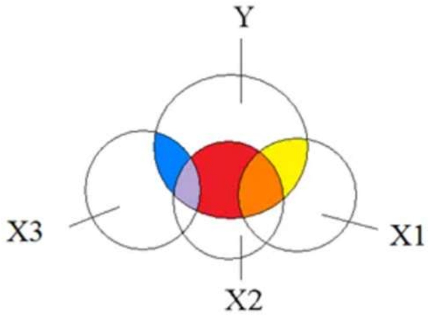


Why is 'high' correlation an issue for model building? When interpreting the model, we make an assumption that the predictors are 'independent'.

What is correlation? This is NOT it!! X1, x2 and x3 are statistically independent. No conditional effects – effects the same as in univariate analysis.

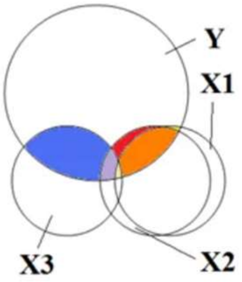
correlation r=0

Relationships between variables - Correlation



Moderate correlation: there is overlap, but conditional effects are still interpretable. **correlation r=0.5**

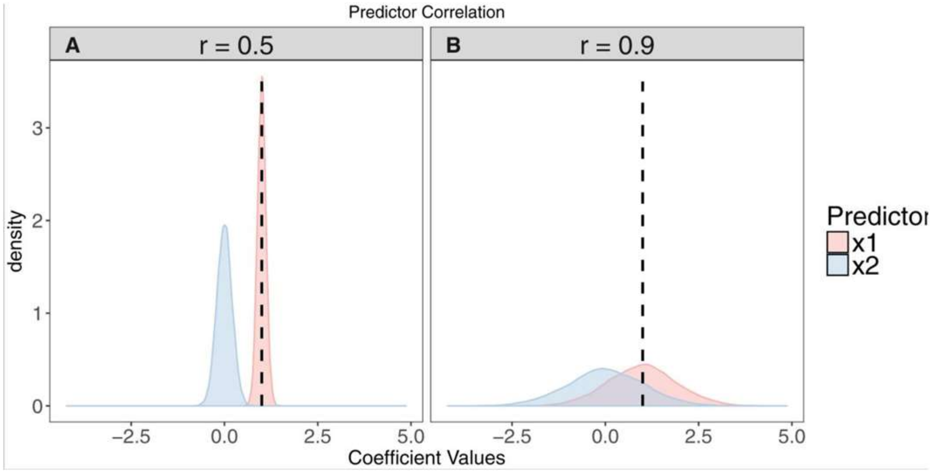
Relationships between variables - Correlation



High correlation: x1 and x2 are almost identical

correlation r=0.9

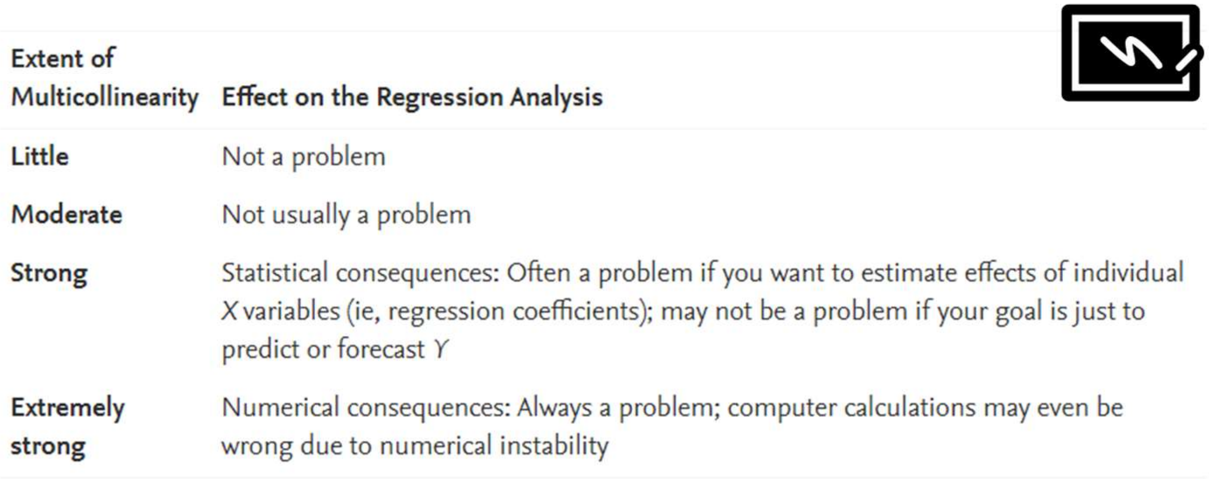
The effect of high correlation: Example: y ~ x1 + x2



Estimate + 95% CI

Moderate correlation **High correlation**

Effect of high correlation/multi-collinearity on modelling



Extent of Multicollinearity	Effect on the Regression Analysis
Little	Not a problem
Moderate	Not usually a problem
Strong	Statistical consequences: Often a problem if you want to estimate effects of individual *X* variables (ie, regression coefficients); may not be a problem if your goal is just to predict or forecast *Y*
Extremely strong	Numerical consequences: Always a problem; computer calculations may even be wrong due to numerical instability

Potential indicators of multi-collinearity:

- Coefficients have signs opposite to what you'd expect from theory (suppression – see DAG in appendix for explanation)
- Very high standard errors for regression coefficients
- Overall model is significant, but none of the coefficients are
- Large changes in coefficients when adding predictors
- Coefficients on different samples are wildly different
- High Variance Inflation Factor (VIF) and low tolerance (VIF reciprocal) – direct measure of how much the variance of the coefficient is being inflated due to multicollinearity – e.g. linear combinations of variables correlated with a variable
- High condition index in PCA – ratio between first and last PC

For further information see: https://www.theanalysisfactor.com/eight-ways-to-detect-multicollinearity/

Step 4: Pick predictors to fit using Exploratory Data Analysis (EDA)

4.2 Assess the relationships among predictors and consider variable reduction

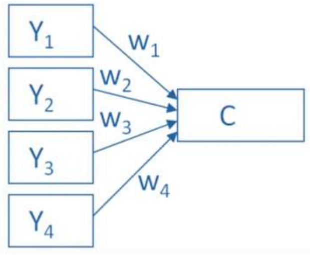
High pairwise correlations – what to do?

- If few variables: select only one of the pair based on biological plausibility, fewer missing values, ease/reliability of measurement or lower univariate p value (may not fix problems arising from collinearity among linear combinations of predictors --> also check Variance Inflation Factors (VIF) after analysis)
- If many (related?) predictors: explore their relationship/consider reducing the number of variables by using multivariate techniques - see our Multivariate Statistical Analysis 1 workshop
- Summarise the variables by creating an index/scale

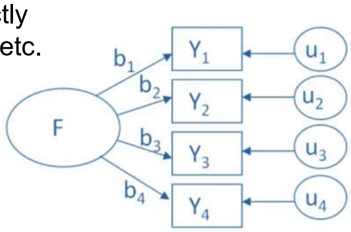
Step 4: Multivariate analysis - Principal component Analysis + Factor Analysis

Principal component Analysis (PCA) and Factor Analysis (FA) are data reduction techniques to consolidate the information contained in a set of numeric predictor variables into a new set of fewer, uncorrelated variables. If used subsequently in a model one cannot statistically test individual predictors.

PCA:
Creates one or more index variables (components) from a larger set of variables by using linear combinations (basically a weighted average) of a set of variables.
Component coefficients from a subsequent model can be back-transformed into coefficients for the original predictors – these are more stable as multi-collinearity is avoided.



FA:
-Models the measurement of a latent variable which cannot be directly measured with a single variable, e.g. intelligence, statistical anxiety etc.
-For subsequent modelling, determining which original predictor is important is subjective based on high correlations/factor loadings.

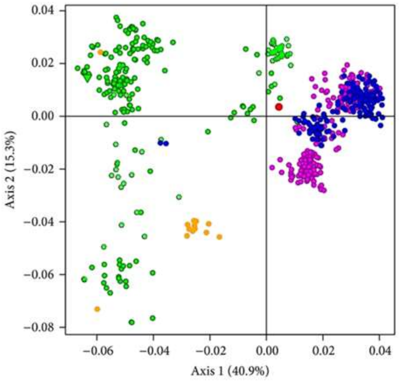


See SIH Multivariate Statistical analysis 1 workshop.

Step 4: Multivariate analysis - Correspondence analysis

- A form of exploratory data analysis (EDA) designed to analyse the relationships among a set of categorical predictor and outcome variables
- Produces a visual summary which is a scatter plot with factorial axes that reflect the most variability in the original predictor variables
- Allows identification of clusters of predictors that are closely associated, with clusters further from the intersection of the axes having stronger associations

Summary:
PCA/FA and correspondence analysis are complementary techniques to modelling and provide insight into relationships between groups of predictors and there association with the outcome



Step 4: Create an index or scale

Combine a number of related predictor variables into a single index

- Subjectively, ideally based on prior research - can have different weights for different contributing factors, e.g. an infection control index may be created by counting the number of hygiene practices performed and expressing this as a percentage of all practices. This could also be categorised into low/medium/high.
- Objectively, e.g. fan capacity, size and number of air inlets and building size may be used to calculate the number of air changes per hour which could also be expressed as proportion of recommended ventilation level
- When predictors are assumed to be reflective of an underlying, unmeasured characteristic (a latent variable) can combine into index or scale by summing or averaging predictors and use Cronbach's alpha statistic to evaluate internal consistency of the scale. Use correlation analysis to assess correlation between each item (variable) and the scale and pairs of items to identify any items that do not fit into the scale well.
  - See SIH workshop on **Surveys' 2** for further information

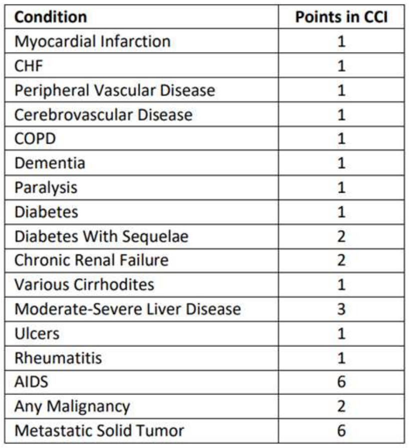
Step 4: Create an index or scale - Comorbidity

Definition:

Comorbidity is defined as the co-occurrence of one or more disorders in the same patient either at the same time or in some causal sequence and a common confounder

- Indices are calculated (weighted sum) from international ICD-10 diagnoses codes (+/- other data) from administrative data (e.g. admitted patients data)
- Indices are based on Elixhauser, the Charlson/Deyo, and the Charlson/Romano methods and well-established for risk adjustment and mortality prediction
- Calculation in R software: **comorbidity package – Rdocumentation**

Charlson comorbidity index



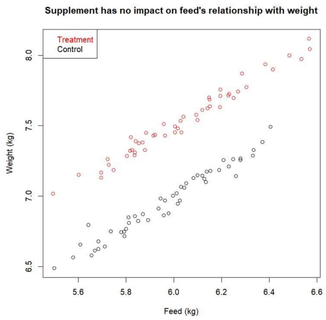
Condition	Points in CCI
Myocardial Infarction	1
CHF	1
Peripheral Vascular Disease	1
Cerebrovascular Disease	1
COPD	1
Dementia	1
Paralysis	1
Diabetes	1
Diabetes With Sequelae	2
Chronic Renal Failure	2
Various Cirrhodites	1
Moderate-Severe Liver Disease	3
Ulcers	1
Rheumatitis	1
AIDS	6
Any Malignancy	2
Metastatic Solid Tumor	6

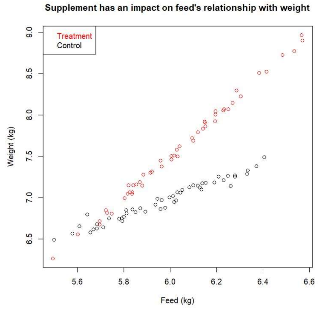
Step 4: Pick a suitable model and predictors to fit using EDA

4.2 Plot the data in relation to the research question, e.g. Plot multiple variables in one plot; plot variables over time

Parallel lines: consistent effect

Non parallel lines: inconsistent effect → **add interaction**


Supplement has no impact on feed's relationship with weight

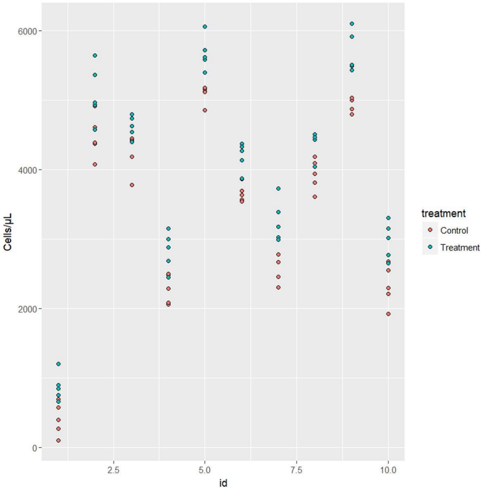

Supplement has an impact on feed's relationship with weight

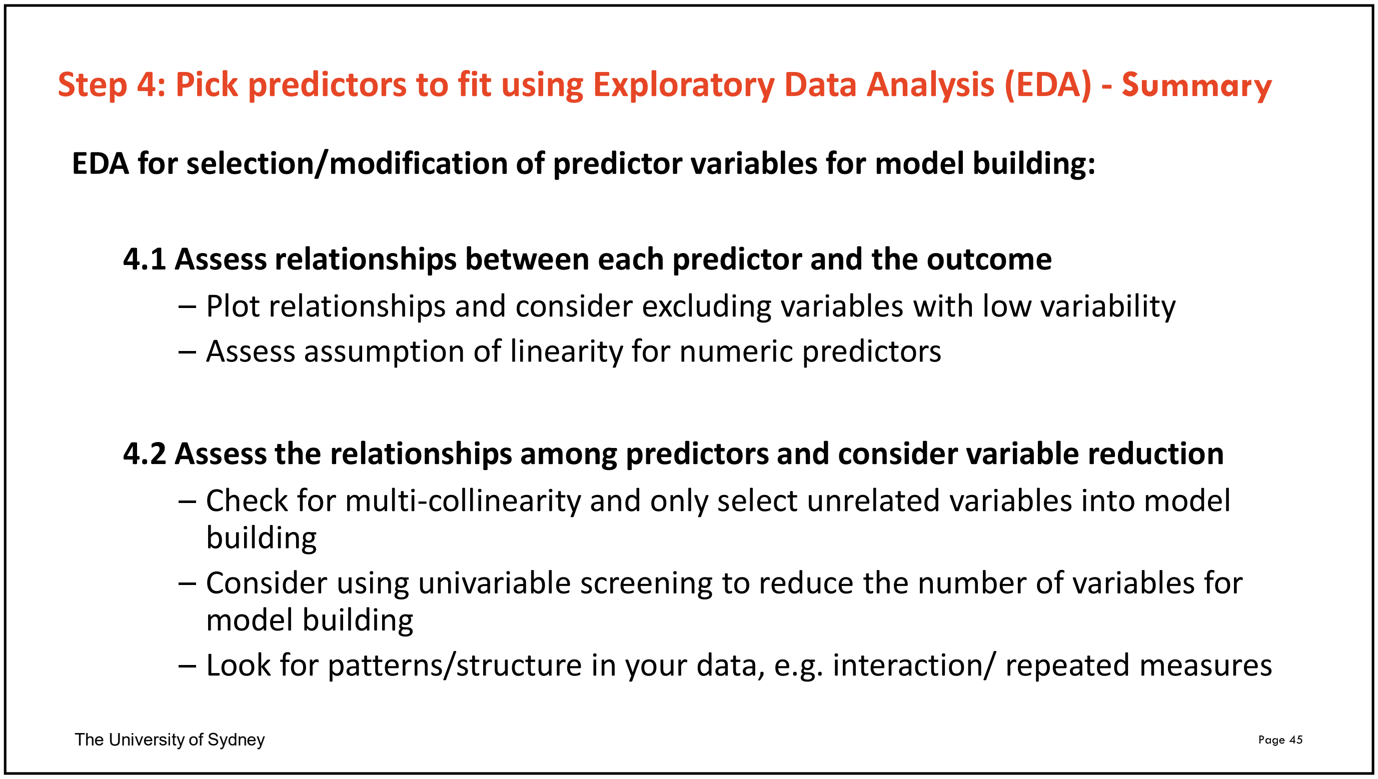
Step 4: Pick a suitable model and predictors to fit using EDA

Plot data for individuals:

The difference between the 10 people (id) is much bigger than the effect of treatment.

→ To account for the variation between people we use a **random effect** for person in a mixed model – this makes our model more accurate and the estimate of treatment effect more precise.





Step 4: Pick predictors to fit using Exploratory Data Analysis (EDA) - Summary

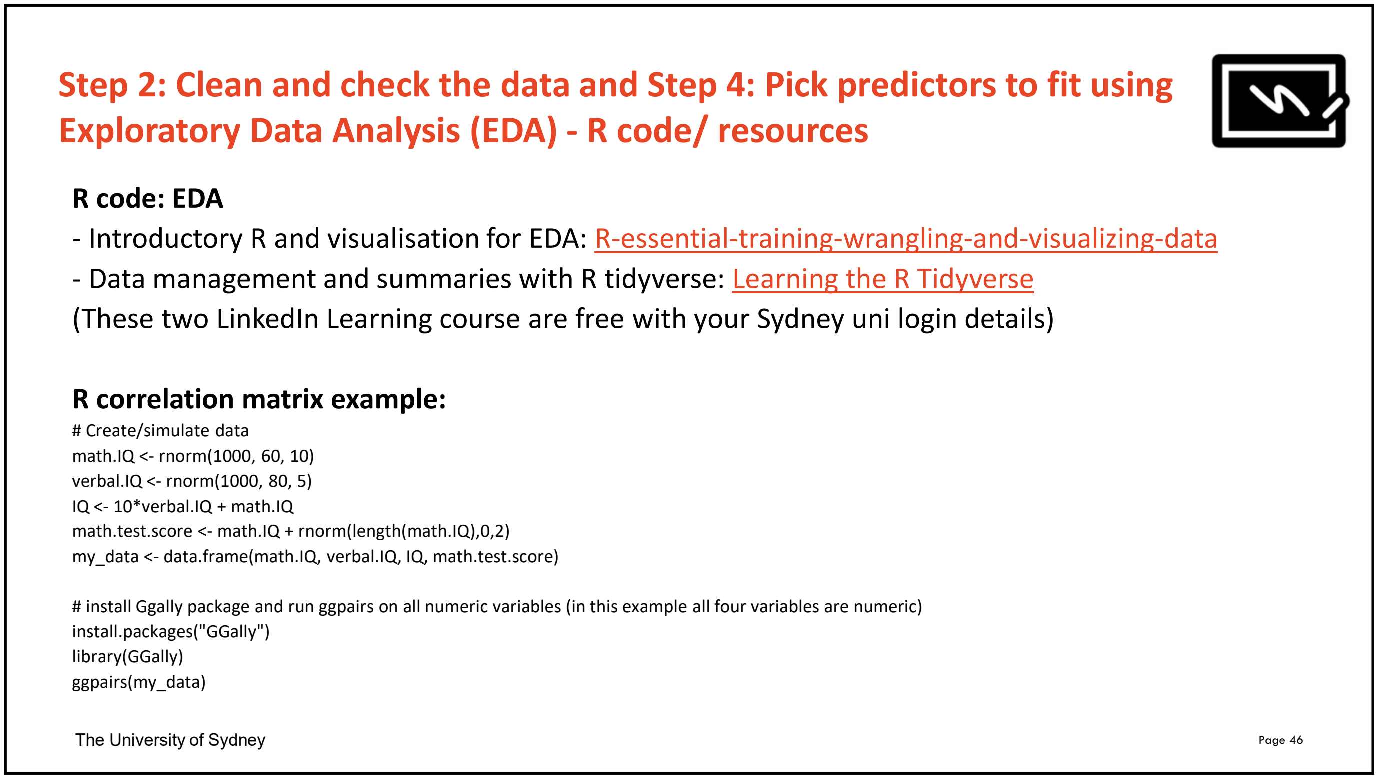
EDA for selection/modification of predictor variables for model building:

4.1 Assess relationships between each predictor and the outcome

- Plot relationships and consider excluding variables with low variability
- Assess assumption of linearity for numeric predictors

4.2 Assess the relationships among predictors and consider variable reduction

- Check for multi-collinearity and only select unrelated variables into model building
- Consider using univariable screening to reduce the number of variables for model building
- Look for patterns/structure in your data, e.g. interaction/ repeated measures



Step 2: Clean and check the data and Step 4: Pick predictors to fit using Exploratory Data Analysis (EDA) - R code/ resources

R code: EDA

- Introductory R and visualisation for EDA: R-essential-training-wrangling-and-visualizing-data

- Data management and summaries with R tidyverse: Learning the R Tidyverse

(These two LinkedIn Learning course are free with your Sydney uni login details)

R correlation matrix example:

```
# Create/simulate data
math.IQ <- rnorm(1000, 60, 10)
verbal.IQ <- rnorm(1000, 80, 5)
IQ <- 10*verbal.IQ + math.IQ
math.test.score <- math.IQ + rnorm(length(math.IQ),0,2)
my_data <- data.frame(math.IQ, verbal.IQ, IQ, math.test.score)

# install Ggally package and run ggpairs on all numeric variables (in this example all four variables are numeric)
install.packages("GGally")
library(GGally)
ggpairs(my_data)
```

Quick break, feel free to stretch your legs!
Any questions?

THE UNIVERSITY OF SYDNEY

47

Workflow: Steps in Model Building

1. Identify the outcome variable and a full set of predictor variables to be considered
2. Clean and check data
3. Pick a suitable modelling method
4. Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)
5. Specify the criterion (criteria) to be used in selecting the variables to be included
6. Specify the strategy for applying the criterion (criteria)
7. Fit the model
8. Check the model assumptions
9. Check model goodness-of-fit
10. Interpret and report the results

48

Step 5: Specify the selection criteria

How to decide which variables to keep in the model?

Selection criteria are formal 'Goodness of Fit' criteria and other statistical considerations

Other statistical considerations - retain variables that are:

- a primary predictor of interest
- a priori confounders for the primary predictor of interest
- Shown confounders for the primary predictor of interest (should not be an intervening variable – see appendix)
- A component of an interaction term included in the model (for interpretation)

Step 5: Specify the selection criteria

Formal Goodness of Fit criteria – nested models

Nested models are based on the same set of observations (n) and the predictors in one model are a subset of predictors in the other model

Tests based on nested models to evaluate significance of a predictor:

- Partial F test (linear model)
- Wald test or likelihood ratio test (LRT) (other types of regression such as logistic, Poisson) – Wald test often most convenient, but LRT has best statistical properties and should be used if p value or standard error are questionable

Formal Goodness of Fit criteria– non-nested models

Information Criteria: IC = -2 log-likelihood + alpha*number of parameters

Akaike's Information Criteria (AIC): alpha = 2

Bayesian Information Criteria (BIC) alpha = log n (with some variation)

Step 5: Specify the selection criteria

AIC/BIC:

– Assess overall model so can be used to compare different models
– The smaller the IC, the better the model
– Can be used to compare nested and non-nested models
– Can be used to compare different regression models, e.g. linear vs Poisson
– Can't be used to compare models based on different sets of observations
– Can't be used to compare models with different likelihood computation, e.g. Cox semi-parametric survival model versus Weibull parametric model
– BIC depends on n and it can be unclear what n to use for clustered data
– BIC favours most parsimonious model

Absolute difference in AIC	Absolute difference in BIC	Evidence for superiority of the better model
0- <4	0 - <2	Weak
4- <7	2- <6	Positive
7 - <10	6 - <10	Strong
=>10	=> 10	Very strong

Step 5: Specify the selection criteria

Two additional approaches for linear regression models:

R squared - amount of variance explained by a univariable model

Adjusted R squared = amount of variance explained by a multi-variable model, penalised for model complexity/number of predictors (larger is better)

- Avoids including predictors that explain a small amount of variance only
- Maximising R squared, minimises mean square error (MSE)

Mallow's Cp Statistic (special case of AIC) :

Cp = Sum((Y-Y_hat)^2)/sigma^2) – n + 2k

Workflow: Steps in Model Building

1. **Identify the outcome variable and a full set of predictor variables to be considered**
2. **Clean and check data**
3. **Pick a suitable modelling method**
4. **Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)**
5. **Specify the criterion (criteria) to be used in selecting the variables to be included**
6. **Specify the strategy for applying the criterion (criteria)**
7. **Fit the model**
8. **Check the model assumptions**
9. **Check model goodness-of-fit**
10. **Interpret and report the results**

Step 6: Specifying the selection strategy

How to get from a full/maximal model (all predictors selected for model building) to the final model?

→Choose a selection strategy to build a model

- Manual options
- Automated processes
- Other considerations

Step 6: Specifying the selection processes – the options

All possible:

Examine all possible combinations of predictors.

Best subset:

Software identifies ‘best’ model of all possible based on criteria (e.g. a model with 2 predictors, largest adjusted R squared; a model with 3 predictors, largest adjusted R squared)

Forward selection

First a model with only the intercept is fitted and then each variables is added selectively based on a specified criterion, e.g. having the largest Wald test statistic provided it corresponds to $p<0.05$ (or other chosen significance level). The predictor with the largest Wald test statistic is added first and then the process is repeated and continues until no term meets the entry criterion.

Backward elimination

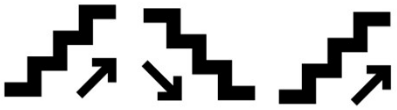
As for forward selection but the process is reversed and model building starts with the full/maximal model. Predictors are removed sequentially until none of the predictors remaining in the model has a Wald test statistic meeting the specified criterion.

Step 6: Specifying the selection process – the options

Stepwise regression

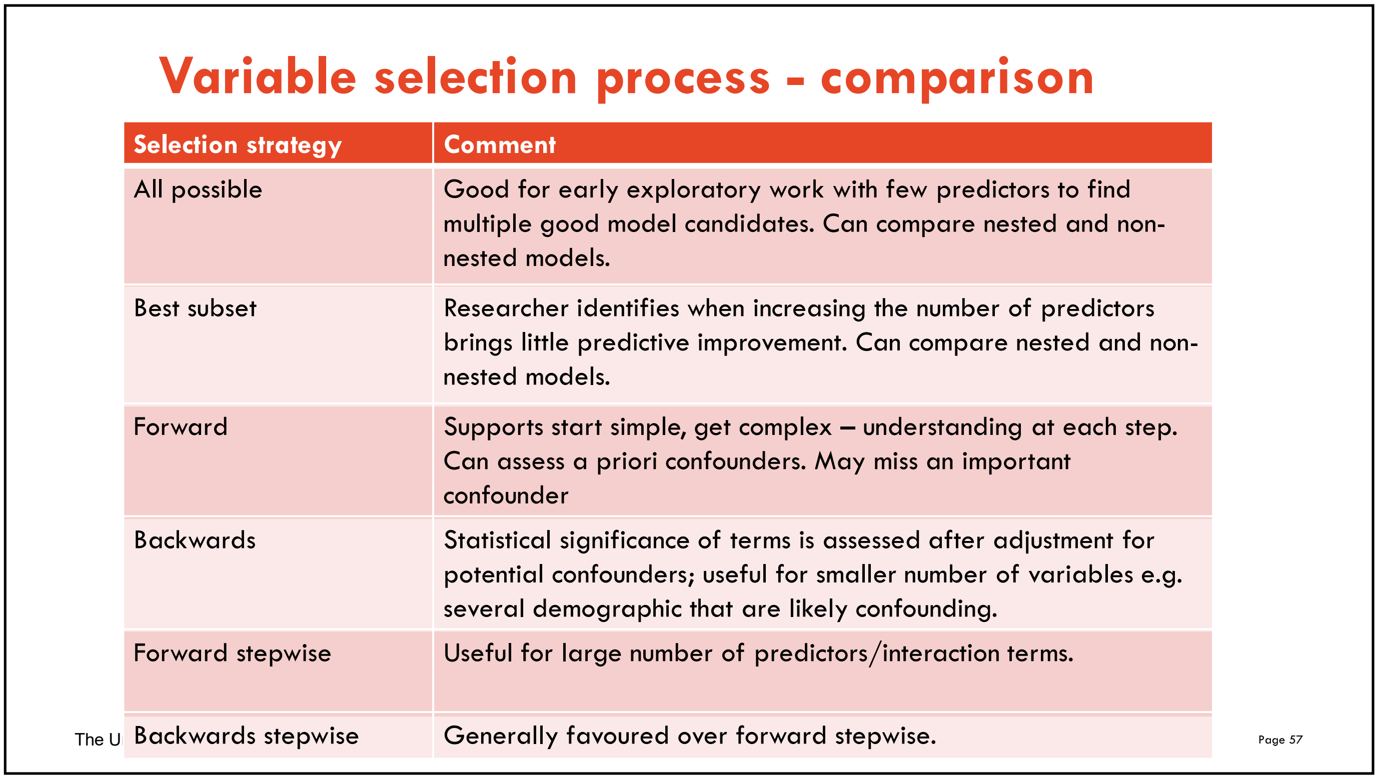
A combination of forward selection and backward elimination.

Forward stepwise starts with forward selection but after the addition of each variable, the criterion for backward elimination is applied to each variable in the model to see if it should remain.



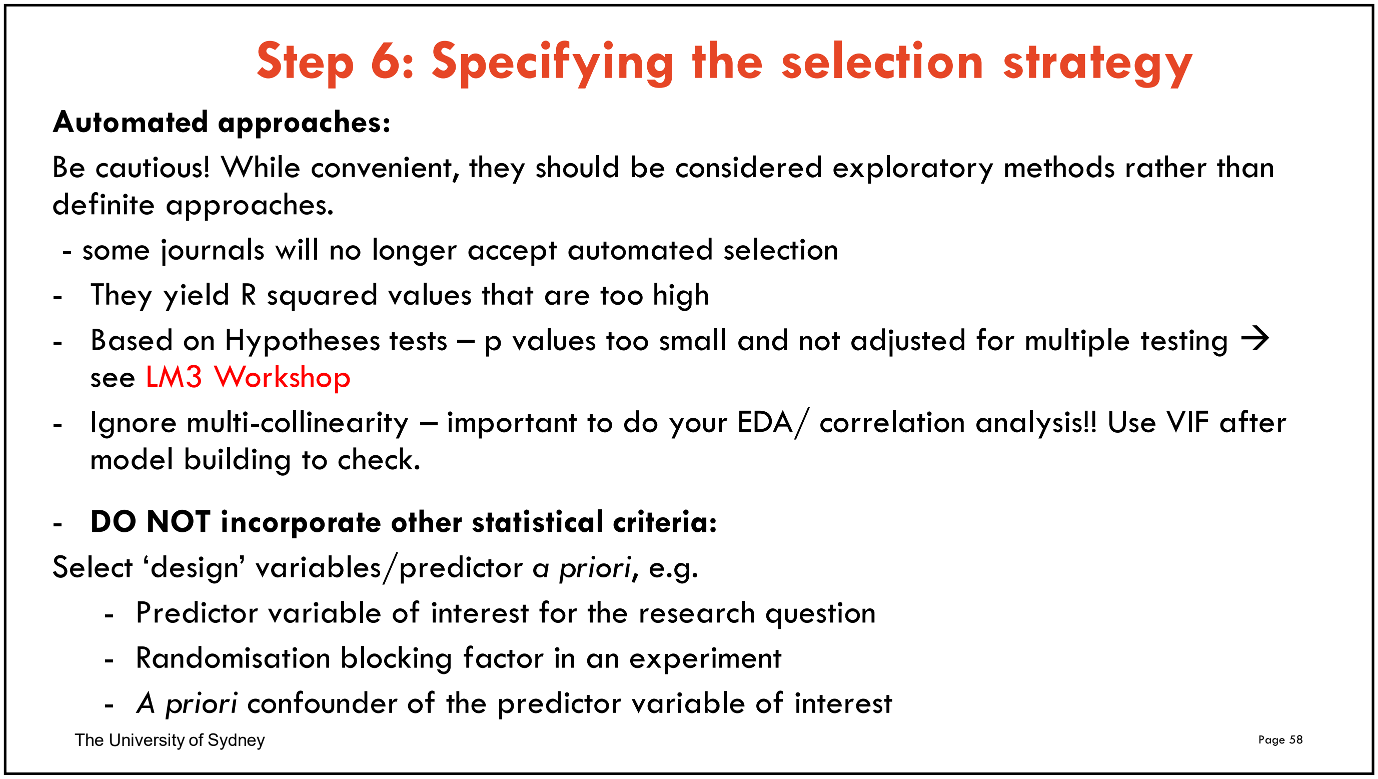
Backward stepwise starts with a full/maximal model and sequentially removes predictors but after the removal of each variable, all removed variables are checked to see if any of them would meet the forward selection criteria for inclusion.

So, what strategy to choose?



Variable selection process - comparison

Selection strategy	Comment
All possible	Good for early exploratory work with few predictors to find multiple good model candidates. Can compare nested and non-nested models.
Best subset	Researcher identifies when increasing the number of predictors brings little predictive improvement. Can compare nested and non-nested models.
Forward	Supports start simple, get complex – understanding at each step. Can assess a priori confounders. May miss an important confounder
Backwards	Statistical significance of terms is assessed after adjustment for potential confounders; useful for smaller number of variables e.g. several demographic that are likely confounding.
Forward stepwise	Useful for large number of predictors/interaction terms.
Backwards stepwise	Generally favoured over forward stepwise.



Step 6: Specifying the selection strategy

Automated approaches:

Be cautious! While convenient, they should be considered exploratory methods rather than definite approaches.

- some journals will no longer accept automated selection
- They yield R squared values that are too high
- Based on Hypotheses tests – p values too small and not adjusted for multiple testing → see LM3 Workshop
- Ignore multi-collinearity – important to do your EDA/ correlation analysis!! Use VIF after model building to check.

- **DO NOT incorporate other statistical criteria:**

Select 'design' variables/predictor *a priori*, e.g.

- Predictor variable of interest for the research question
- Randomisation blocking factor in an experiment
- *A priori* confounder of the predictor variable of interest

Model building strategy styles (adapted from Keith McCormick)

	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

Source: LinkedIn Learning video: Three regression strategies, Keith McCormick.

Model Building Strategy - 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.

Pitfalls with automated selection processes:

Avoids researcher thinking about their data and taking responsibility for their analysis!

"Start simple, get complex"

"Every model you run tells you a story. Stop and listen to it."

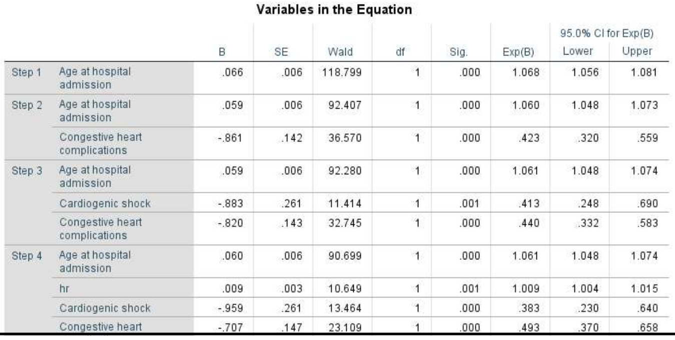
- Avoid Overfitting (have at least 10 (better 20!) observations per parameter in the model – keep in mind dummy coded categorical variables with > 2 categories e.g. a predictor with 5 categories requires 4 parameters to be estimated)
- For interpretability of parameters keep the main effects of both variables that make up a significant 2-way interaction term in the model (irrespective of their p values)
- Analysis will only be based on observations for which all variables are not missing (with many missing observations the final data set analysed may be a small subset) – do EDA!

Workflow: Steps in Model Building

1. **Identify the outcome variable and a full set of predictor variables to be considered**
2. **Clean and check data**
3. **Pick a suitable modelling method**
4. **Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)**
5. **Specify the criterion (criteria) to be used in selecting the variables to be included**
6. **Specify the strategy for applying the criterion (criteria)**
7. **Fit the model**
8. **Check the model assumptions**
9. **Check model goodness-of-fit**
10. **Interpret and report the results**

Step 7: Fit the model/conduct the analysis

- **An iterative process**
- Build and evaluate models
- Gain insights into the complex relationships among the variables ➔ allows for more refined, biologically reasonable models to be build
- Incorporate expert knowledge of the system being studied along with the results of the analyses



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	-.707	.147	23.109	1	.000	.493	.370	.658

Workflow: Steps in Model Building

1. **Identify the outcome variable and a full set of predictor variables to be considered**
2. **Clean and check data**
3. **Pick a suitable modelling method**
4. **Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)**
5. **Specify the criterion (criteria) to be used in selecting the variables to be included**
6. **Specify the strategy for applying the criterion (criteria)**
7. **Fit the model**
8. **Check the model assumptions**
9. **Check model goodness-of-fit**
10. **Interpret and report the results**

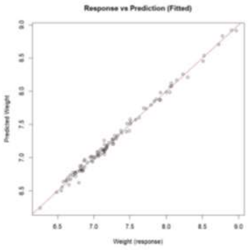
Step 8: Check the model assumptions

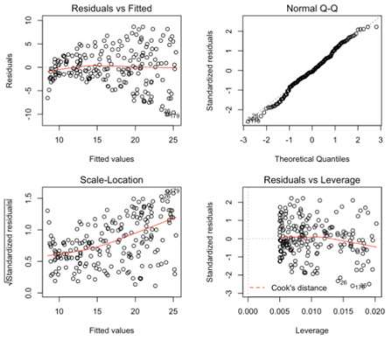
Use “diagnostics” to assess the *validity* of the model, e.g.

- Evaluate normality of residuals in LM
- Check HR/OR proportionality assumptions
- Check for influential values
- Check assumption of independent predictors using Variance Inflation Factors (VIF) for correlation among predictors after analysis:

VIF >4 moderate multi-collinearity

VIF >10 strong multi-collinearity





Workflow: Steps in Model Building

1. **Identify the outcome variable and a full set of predictor variables to be considered**
2. **Clean and check data**
3. **Pick a suitable modelling method**
4. **Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)**
5. **Specify the criterion (criteria) to be used in selecting the variables to be included**
6. **Specify the strategy for applying the criterion (criteria)**
7. **Fit the model**
8. **Check the model assumptions**
9. **Check model goodness-of-fit**
10. **Interpret and report the results**

Step 9: Check model fit

Evaluate the *reliability* of the model – how well will the model predict observations in future samples?

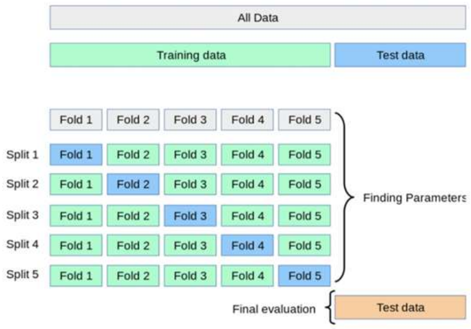
Different approaches, e.g.:

– Split-sample analysis

– Cross-validation

– Leave-one-out analysis

– Bootstrap



K-Fold Cross Validation

– Goodness-of-fit test, e.g. Hosmer-Lemeshow test

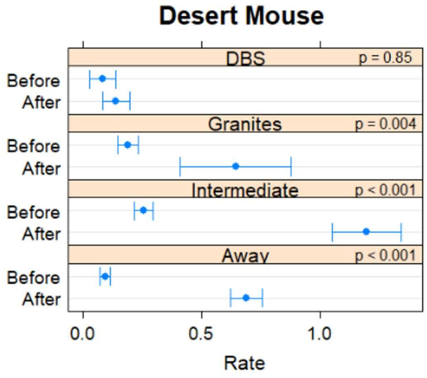
Workflow: Steps in Model Building

1. **Identify the outcome variable and a full set of predictor variables to be considered**
2. **Clean and check data**
3. **Pick a suitable modelling method**
4. **Pick predictors to fit and a suitable model using Exploratory Data Analysis (EDA)**
5. **Specify the criterion (criteria) to be used in selecting the variables to be included**
6. **Specify the strategy for applying the criterion (criteria)**
7. **Fit the model**
8. **Check the model assumptions**
9. **Check model goodness-of-fit**
10. **Interpret and report the results**

Step 10: Interpret and report the results

- Present point estimates – the coefficients (including the intercept)
- Present the measure of uncertainty – their standard errors and/or their confidence interval.

Depending on target audience, non-numerical presentation of study results may be preferable.



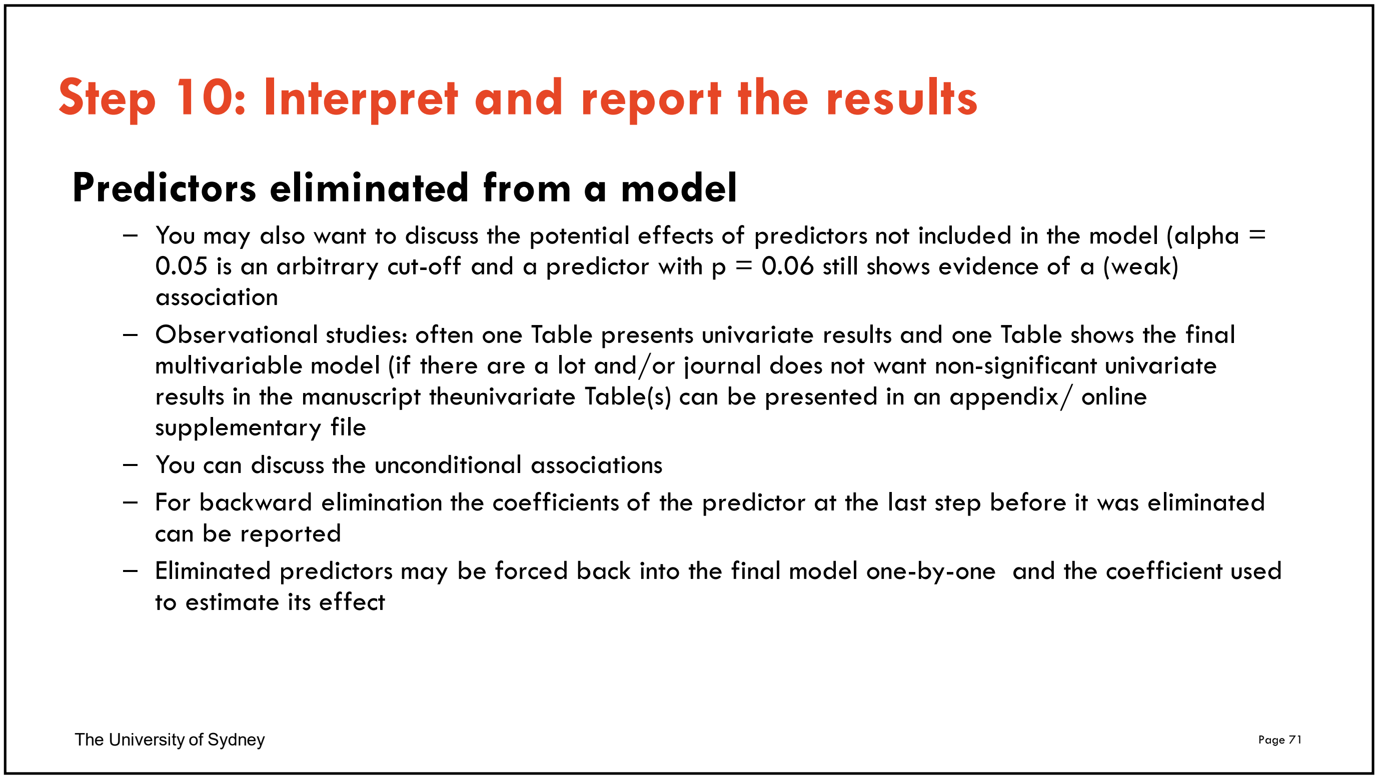
Step 10: Interpret and report the results – example results tables

Table 1: Univariable linear regression model results showing the association between offspring height with father's height, mother's height, gender and financial situation in a study of 288 university students conducted in Sydney.

Variable		$\hat{b}$	S.E.($\hat{b}$)	95% CI	P value	Percentage variance accounted for (Adjusted R^2)
Constant		90.0	12.3	(65.9, 114.1)	<.001	13.7
Father height		0.47	0.07	(0.33, 0.60)		
Constant		71.4	14.3	(43.2, 99.6)	<.001	15.1
Mother height		0.62	0.09	(0.45, 0.79)		
Gender (reference=Female)		167.0	0.56	(165.9, 168.1)	<.001	46.8
	Male	13.98	0.88	(12.2, 15.7)		
Finance (reference=Finding it difficult)		171.5	1.17	(169.2, 173.8)	0.147	0.8
	Just getting along	0.33	1.69	(-3.00, 3.66)		
	Comfortable	0.61	1.59	(-2.52, 3.74)		
	Prosperous	3.63	1.73	(0.23, 7.03)		

Table 2: Final multivariable linear regression model for offspring height in a study of 288 university students in Sydney

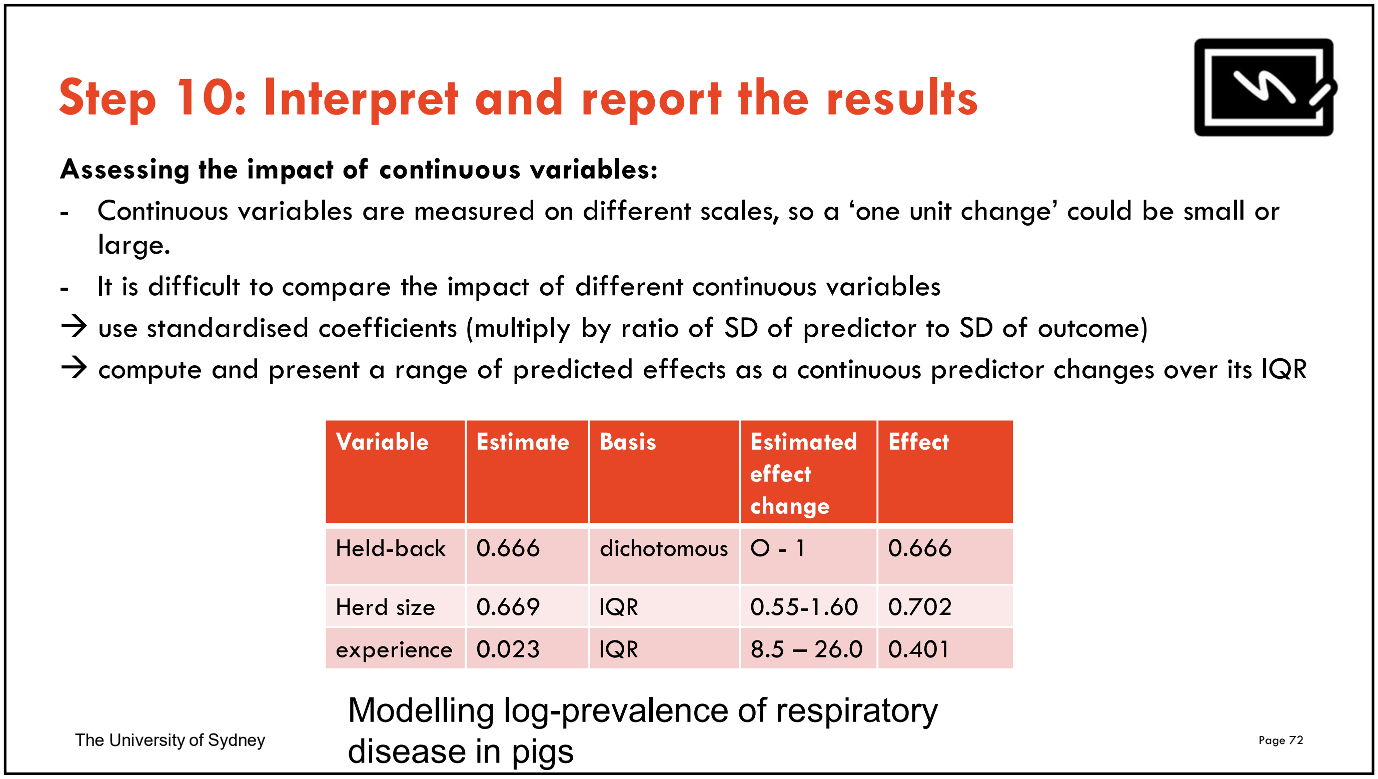
Variable		Estimate	S.E.	95% CI		P value
Intercept		24.70	10.20	4.71	44.69	0.016
Mother_ht		0.41	0.05	0.31	0.52	<.001
Father_ht		0.42	0.04	0.34	0.51	<.001
Gender	Male versus Female	14.19	0.69	12.84	15.53	<.001



Step 10: Interpret and report the results

Predictors eliminated from a model

- You may also want to discuss the potential effects of predictors not included in the model (alpha = 0.05 is an arbitrary cut-off and a predictor with p = 0.06 still shows evidence of a (weak) association
- Observational studies: often one Table presents univariate results and one Table shows the final multivariable model (if there are a lot and/or journal does not want non-significant univariate results in the manuscript theunivariate Table(s) can be presented in an appendix/ online supplementary file
- You can discuss the unconditional associations
- For backward elimination the coefficients of the predictor at the last step before it was eliminated can be reported
- Eliminated predictors may be forced back into the final model one-by-one and the coefficient used to estimate its effect



Step 10: Interpret and report the results

Assessing the impact of continuous variables:

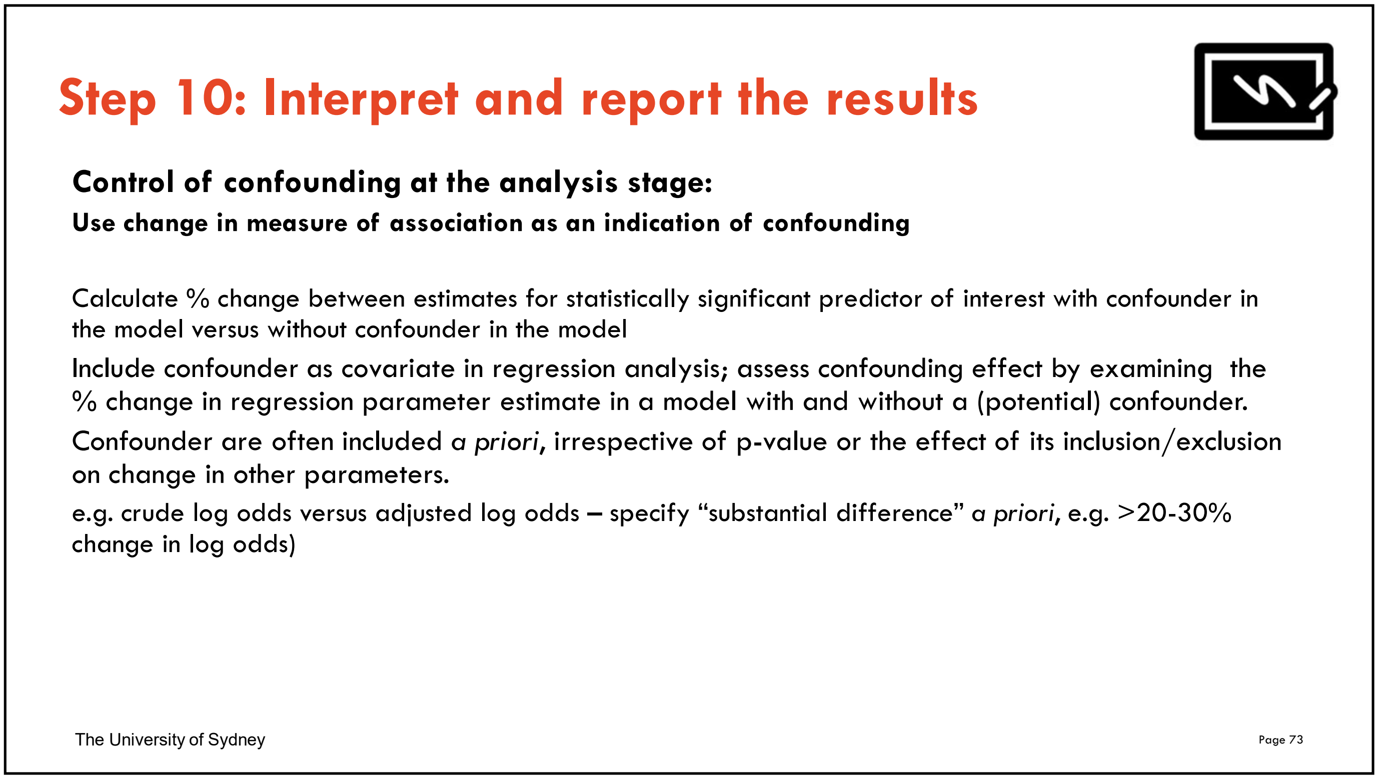
- Continuous variables are measured on different scales, so a 'one unit change' could be small or large.
- It is difficult to compare the impact of different continuous variables

→ use standardised coefficients (multiply by ratio of SD of predictor to SD of outcome)

→ compute and present a range of predicted effects as a continuous predictor changes over its IQR

Variable	Estimate	Basis	Estimated effect change	Effect
Held-back	0.666	dichotomous	O - 1	0.666
Herd size	0.669	IQR	0.55-1.60	0.702
experience	0.023	IQR	8.5 – 26.0	0.401

Modelling log-prevalence of respiratory disease in pigs



Step 10: Interpret and report the results

Control of confounding at the analysis stage:

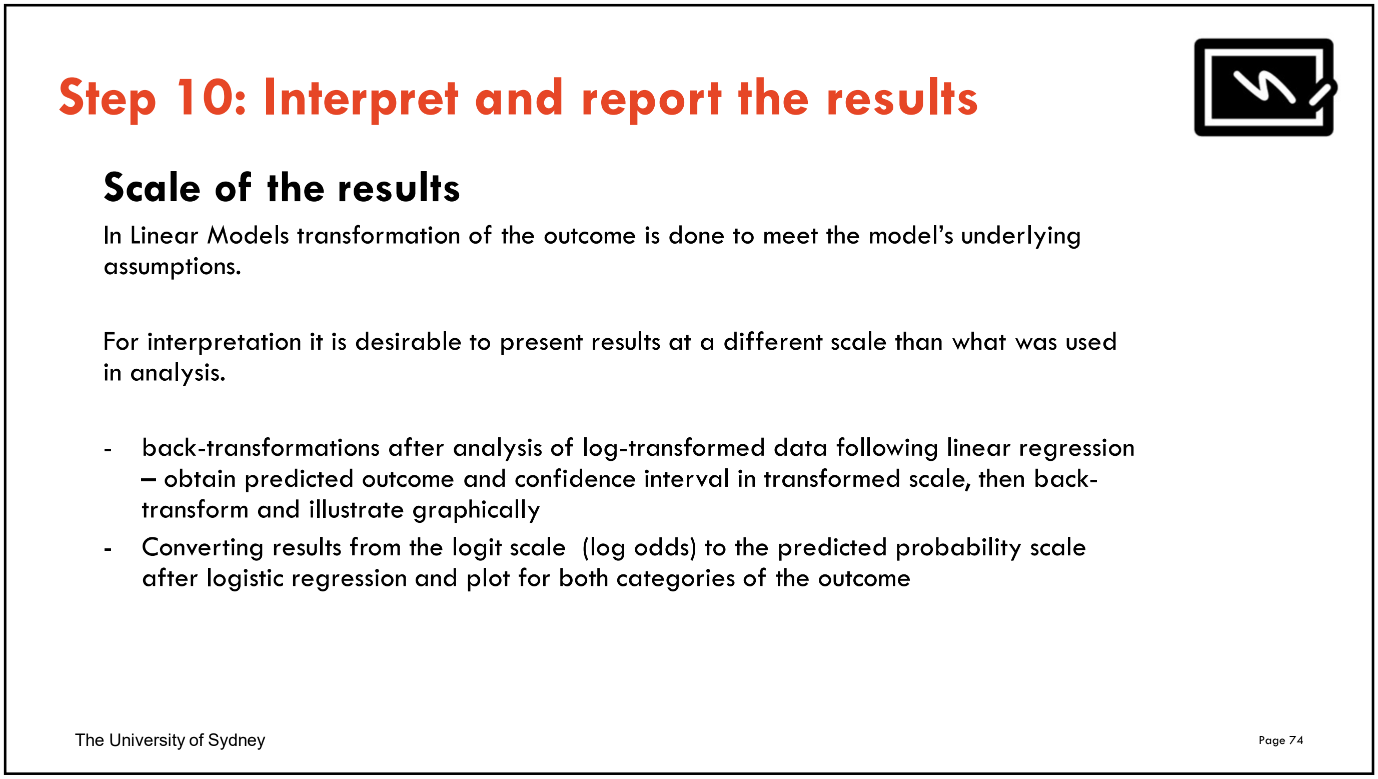
Use change in measure of association as an indication of confounding

Calculate % change between estimates for statistically significant predictor of interest with confounder in the model versus without confounder in the model

Include confounder as covariate in regression analysis; assess confounding effect by examining the % change in regression parameter estimate in a model with and without a (potential) confounder.

Confounder are often included *a priori*, irrespective of p-value or the effect of its inclusion/exclusion on change in other parameters.

e.g. crude log odds versus adjusted log odds – specify "substantial difference" *a priori*, e.g. >20-30% change in log odds)



Step 10: Interpret and report the results

Scale of the results

In Linear Models transformation of the outcome is done to meet the model's underlying assumptions.

For interpretation it is desirable to present results at a different scale than what was used in analysis.

- back-transformations after analysis of log-transformed data following linear regression – obtain predicted outcome and confidence interval in transformed scale, then back-transform and illustrate graphically
- Converting results from the logit scale (log odds) to the predicted probability scale after logistic regression and plot for both categories of the outcome

Model Building: R software resources

Linear model building example in R and other relevant resources for R:
https://r4ds.had.co.nz/model-building.html

Questions?

Upcoming SIH statistical trainings

Check out the SIH training calendar or sign up to the mailing list for details on all upcoming trainings:

https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training/training-calendar.html

Consider our monthly informal "Hacky Hour" session for all your data management, visualization and coding queries:

Sydney Hacky Hour - The University of Sydney

Further Assistance at Sydney University

SIH

- **Workshops**

This is just one workshop from a modular training programme made up of 1.5 hour workshops, each focusing on a single statistical method offered by Statistical Consulting within the Sydney Informatics Hub. Statistical Workflows giving practical step-by-step instructions applicable in any software are used and include experimental design, exploratory analysis, modelling, assumption testing, model interpretation and presentation of results. They are integrated into Training Pathways (insert link) to give a holistic understanding of data analysis from a statistical perspective. Researchers are also encouraged to design a custom programme tailored to their research needs.

Look for the statistics workshops (and other SIH workshops) on our training page https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training.html#stats.

- **Statistical Resources from the Sydney Informatics Hub | stats-resources (sydney-informatics-hub.github.io)**
- **Training: Sign up to our mailing list to be notified of upcoming training: mailman.sydney.edu.au/mailman/listinfo/computing_training**
- **Hacky Hour** www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training/hacky-hour.html **OR Google "Sydney Hacky Hour"**
- **1on1 Consults can be requested on our website** www.sydney.edu.au/research/facilities/sydney-informatics-hub.html **OR Google "Sydney Informatics Hub"**

OTHER

- **Open Learning Environment (OLE) courses**
- **Linkedin Learning:** https://linkedin.com/learning/
  - **SPSS** https://www.linkedin.com/learning/machine-learning-ai-foundations-linear-regression/welcome?u=2196204

We recommend our Experimental Design and Sample Size Workshops

Experimental Design Workshop

- Far too many researchers think they know all they need to in this area. We commonly see designs that could be substantially improved for stronger causal inference and improved results which leads to publication in higher impact journals (amongst other benefits).
- Even if you have already collected your data it is well worth attending since it may improve your write up and analysis e.g. we had a client who didn't realise they had a very strong Before/After Control/Impact (BACI) design.

Sample and Power Workshop

- Shows the steps and decisions researchers need to make when designing an experiments to ensure sufficient sample e.g. Power, minimum required to fit the necessary model, etc.
- Also how much Power the study has i.e. does it have sufficient power to detect the effects you expect to see, or is your study a complete waste of time and resources.

A reminder about 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 statistical consulting service provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."
Acknowledging specific staff:
"The authors acknowledge the statistical consulting service provided by (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

"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.**

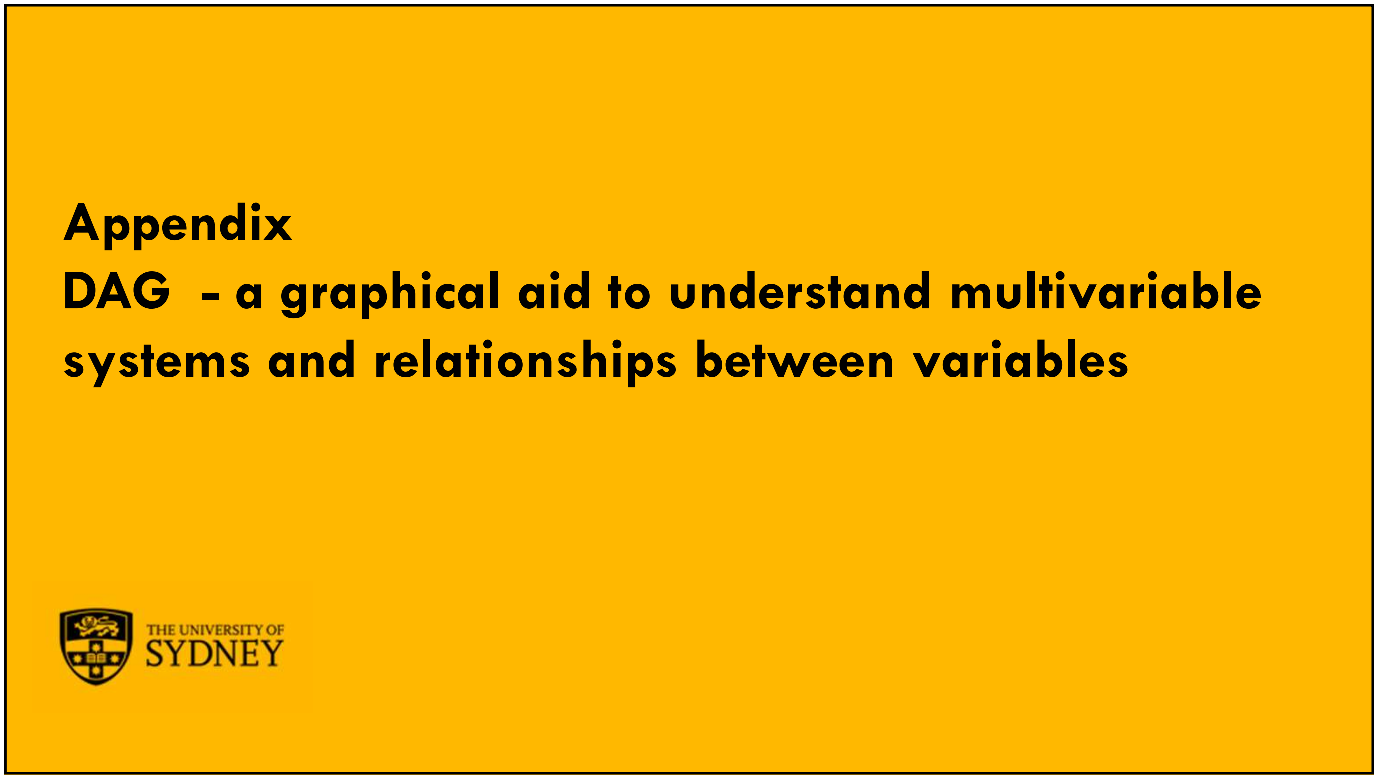
We actively use the feedback to improve our workshops.

Completing this survey really does help us and we would appreciate your help! It only takes a few minutes to complete (*promise!*)

The link to the survey will be emailed.

References used when building this Workshop

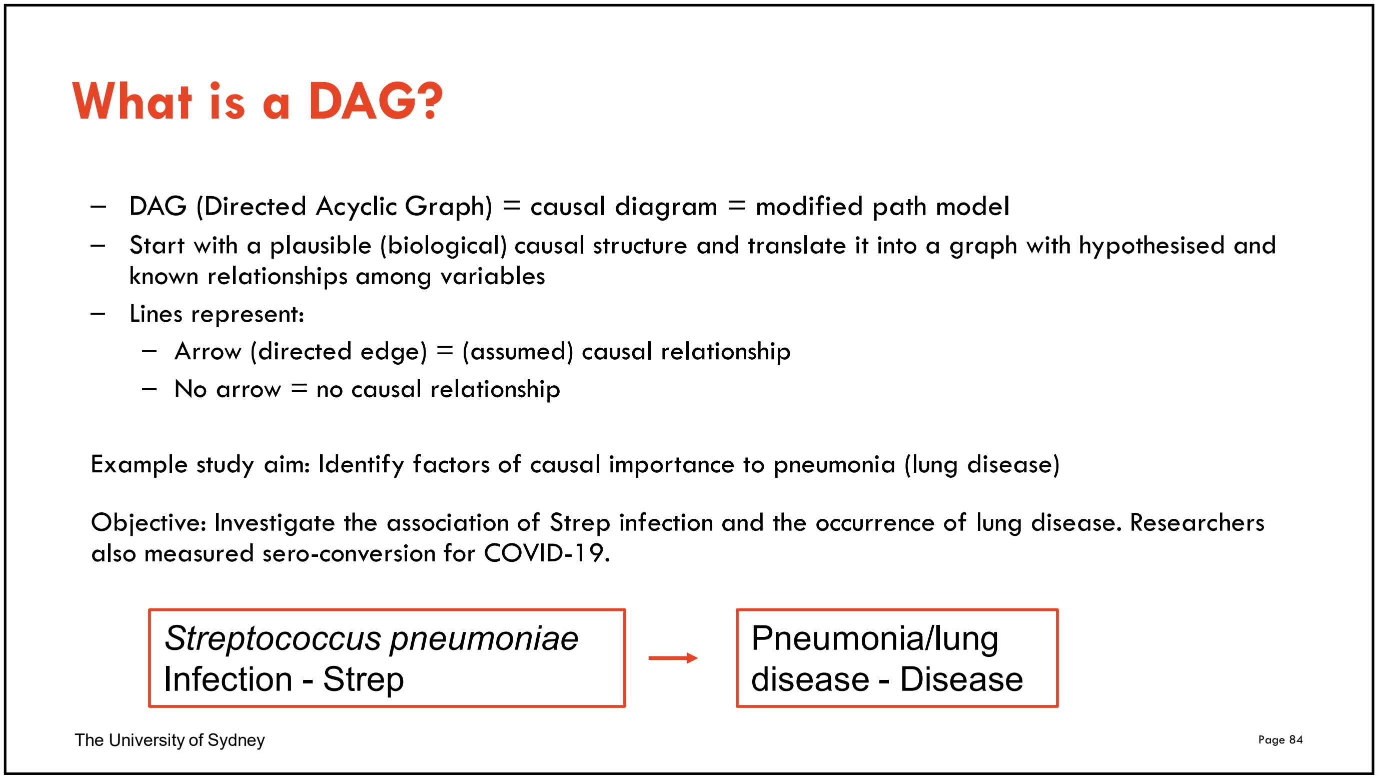
- Dohoo et al 2009 Veterinary Epidemiologic Research
- Rothman et al 2008 Modern Epidemiology, 3rd ed, Chapter 12
- Harrell, Frank E. Regression Modeling Strategies : with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis . Second edition. Cham: Springer, 2015. Print. https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/14vvljs/alma991008621859705106
- Steyerberg EW (2019) Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating https://sydney.primo.exlibrisgroup.com/permalink/61USYD_INST/1367smt/cdi_askewsholts_vlebooks_9783030163990
- Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. *Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data.* Medical Care 2005; 43(11):1130-1139. DOI: 10.1097/01.mlr.0000182534.19832.83



Appendix
DAG - a graphical aid to understand multivariable systems and relationships between variables

THE UNIVERSITY OF SYDNEY

83



What is a DAG?

- DAG (Directed Acyclic Graph) = causal diagram = modified path model
- Start with a plausible (biological) causal structure and translate it into a graph with hypothesised and known relationships among variables
- Lines represent:
  - Arrow (directed edge) = (assumed) causal relationship
  - No arrow = no causal relationship

Example study aim: Identify factors of causal importance to pneumonia (lung disease)

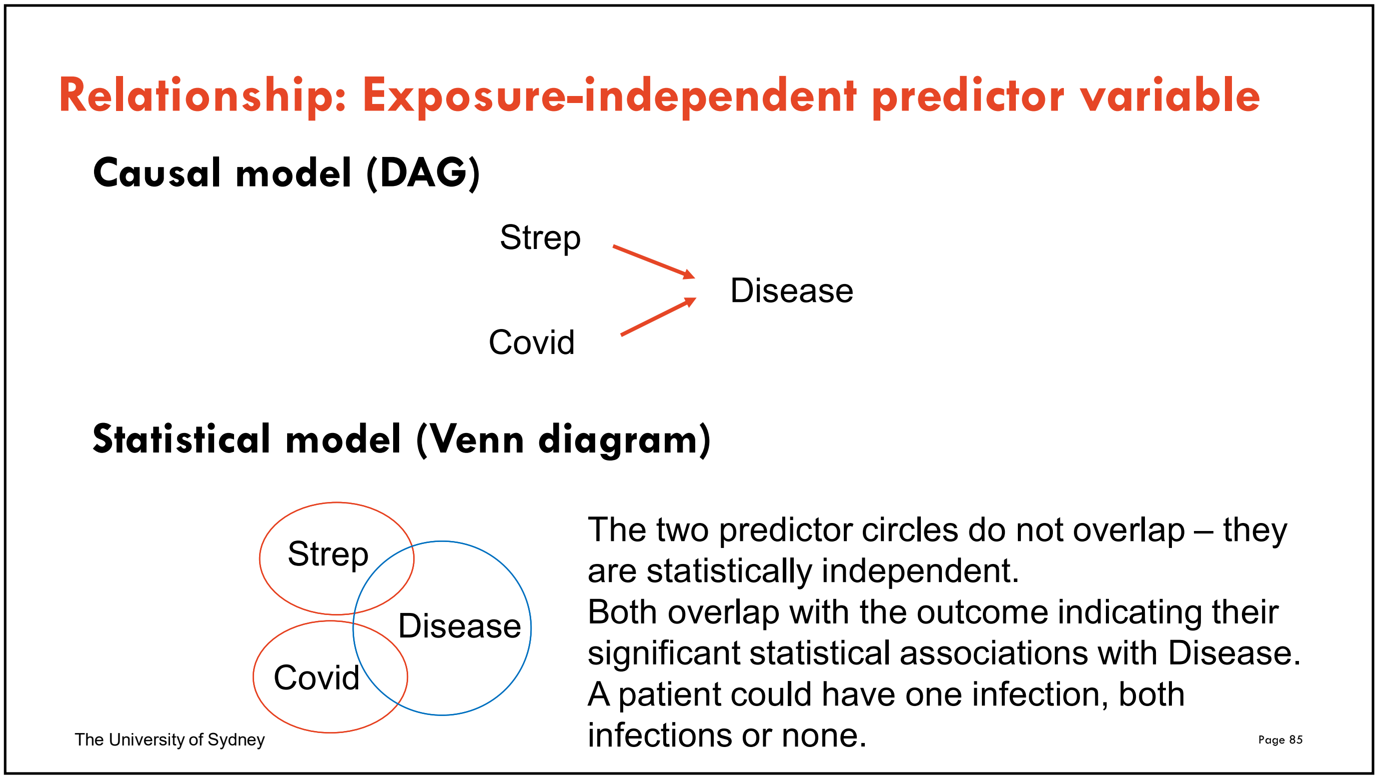
Objective: Investigate the association of Strep infection and the occurrence of lung disease. Researchers also measured sero-conversion for COVID-19.

Streptococcus pneumoniae Infection - Strep → Pneumonia/lung disease - Disease

The University of Sydney

Page 84

84

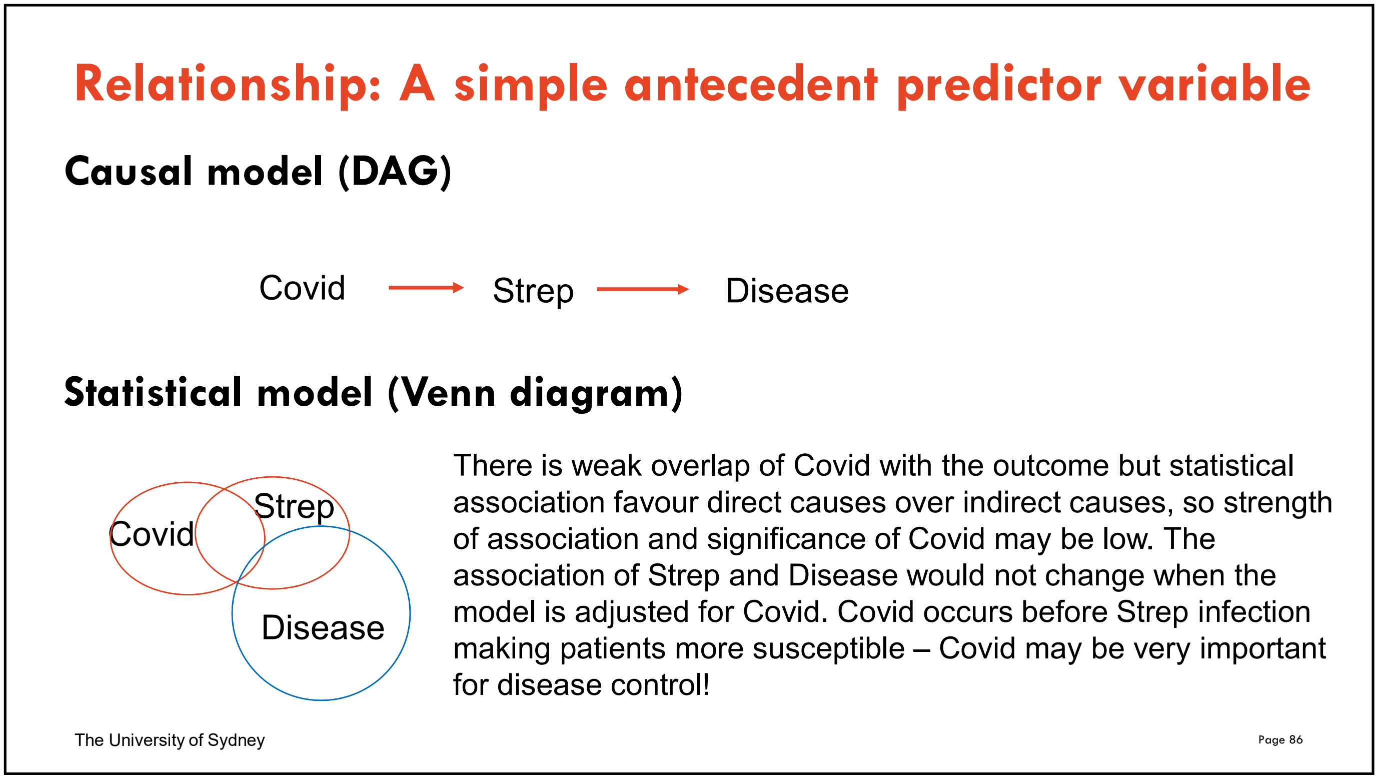


Relationship: Exposure-independent predictor variable

Causal model (DAG)

Statistical model (Venn diagram)

The two predictor circles do not overlap – they are statistically independent.
Both overlap with the outcome indicating their significant statistical associations with Disease.
A patient could have one infection, both infections or none.



Relationship: A simple antecedent predictor variable

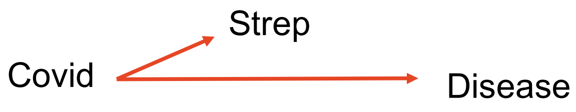
Causal model (DAG)

Statistical model (Venn diagram)

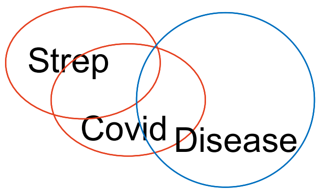
There is weak overlap of Covid with the outcome but statistical association favour direct causes over indirect causes, so strength of association and significance of Covid may be low. The association of Strep and Disease would not change when the model is adjusted for Covid. Covid occurs before Strep infection making patients more susceptible – Covid may be very important for disease control!

Relationship: An explanatory antecedent variable – complete confounding

Causal model (DAG)



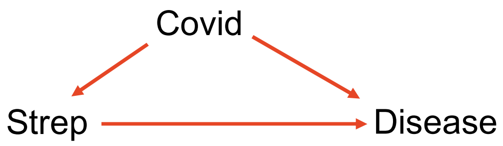
Statistical model (Venn diagram)



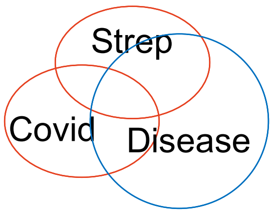
The Strep circle overlaps with the outcome, as they are statistically related until Covid is added to the model. Then the association becomes non-significant as all of the previous crude association is covered by the Covid-Disease association.

Relationship: an explanatory antecedent variable – partial confounding

Causal model (DAG)

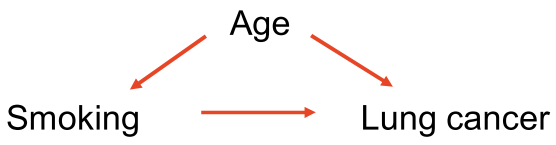


Statistical model (Venn diagram)



The Strep circle overlaps with the outcome. The association remains significant when Covid is added to the model but some of the previous association is now covered by the Covid-Disease association.
The Strep-Disease association is not as strong when Covid confounding is controlled but the model with both predictors explains more variation in Disease than just Strep.

More on confounding: a classic example


Age

Smoking

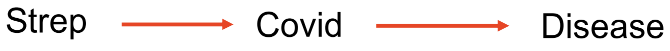
Lung cancer

Age confounds the relationship of smoking and lung cancer – smoking may be more prevalent among older people and older people are more likely to get lung cancer irrespective of smoking.

Examples of common confounders:
Age, Sex, Socio-economic status, Comorbidity, previous experience, etc.

Relationships: an intervening predictor variable

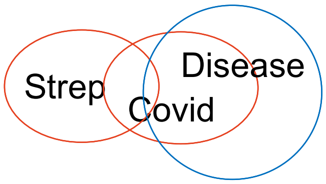
Causal model (DAG)


Strep

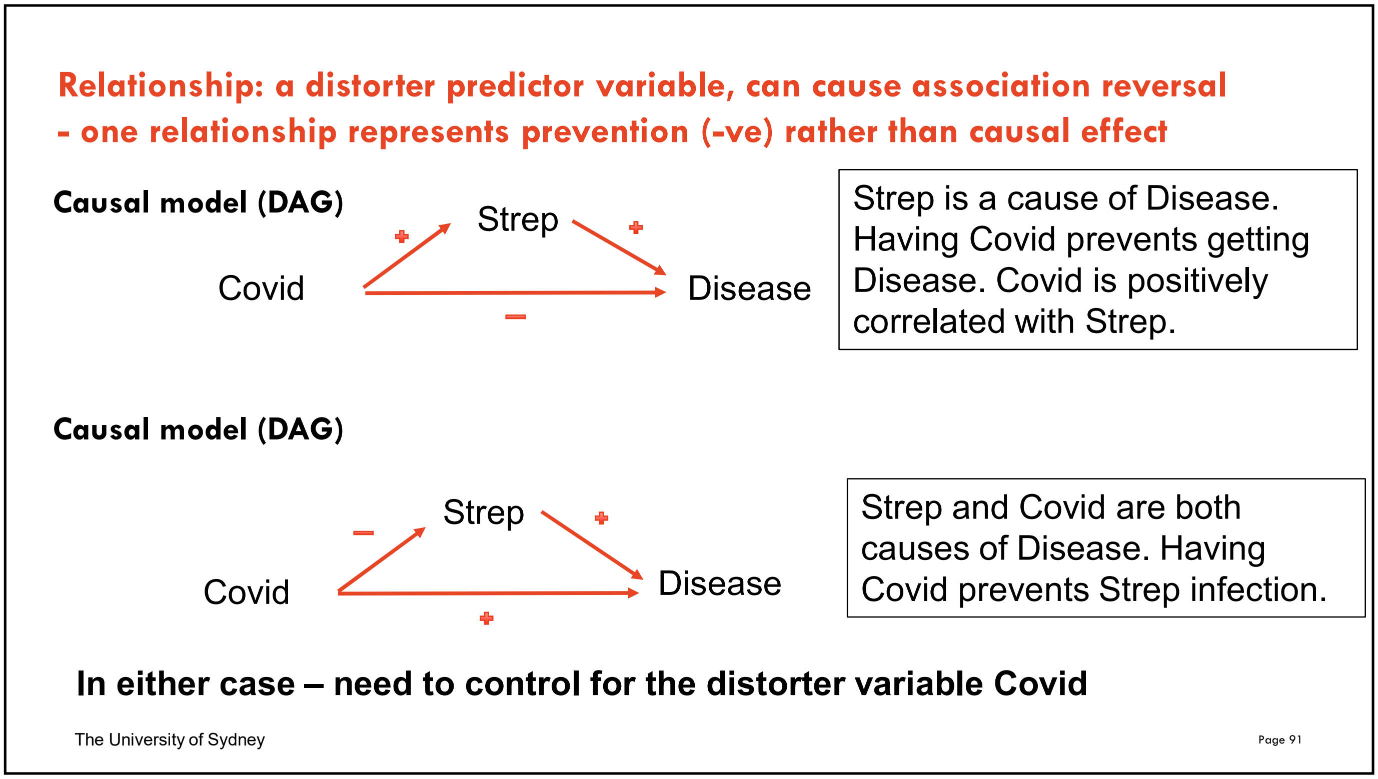
Covid

Disease

Statistical model (Venn diagram)



The Strep circle might or might not overlap with the outcome. However, any association of Strep with Disease disappears when Covid is added to the model. The effect of Strep on Disease is mediated by Covid. Adding Covid to the model would lead us to believe that Strep is not associated with Disease (and hence we may wrongly conclude that Strep is not a cause of Disease). Intervening variables should be identified and not be used/controlled when estimating the causal effect of an exposure → consider Structural equation modelling, mediation analysis instead if interested in the mediator.



Relationship: a distorter predictor variable, can cause association reversal - one relationship represents prevention (-ve) rather than causal effect

Causal model (DAG)

Strep
Covid
Disease
+
+
–

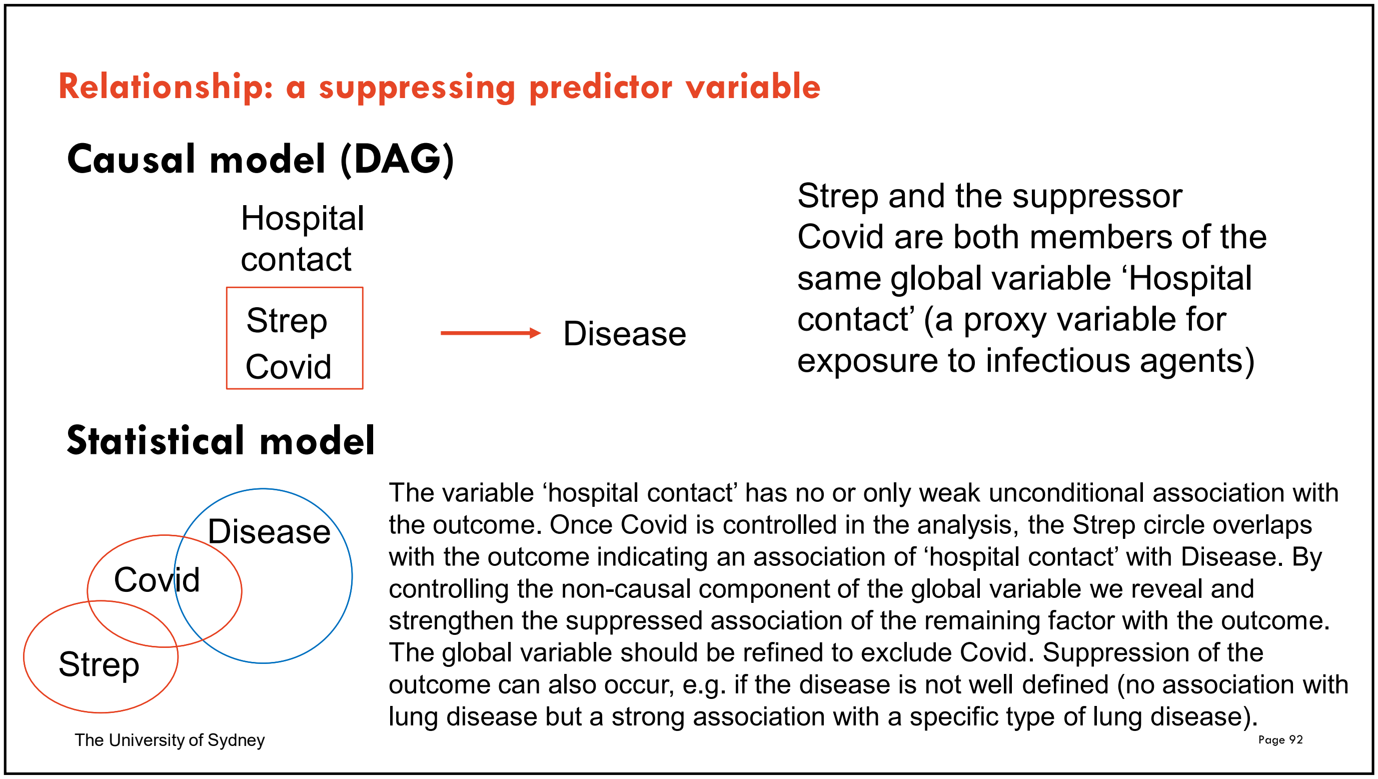
Strep is a cause of Disease. Having Covid prevents getting Disease. Covid is positively correlated with Strep.

Causal model (DAG)

Strep
Covid
Disease
–
+
+

Strep and Covid are both causes of Disease. Having Covid prevents Strep infection.

In either case – need to control for the distorter variable Covid



Relationship: a suppressing predictor variable

Causal model (DAG)

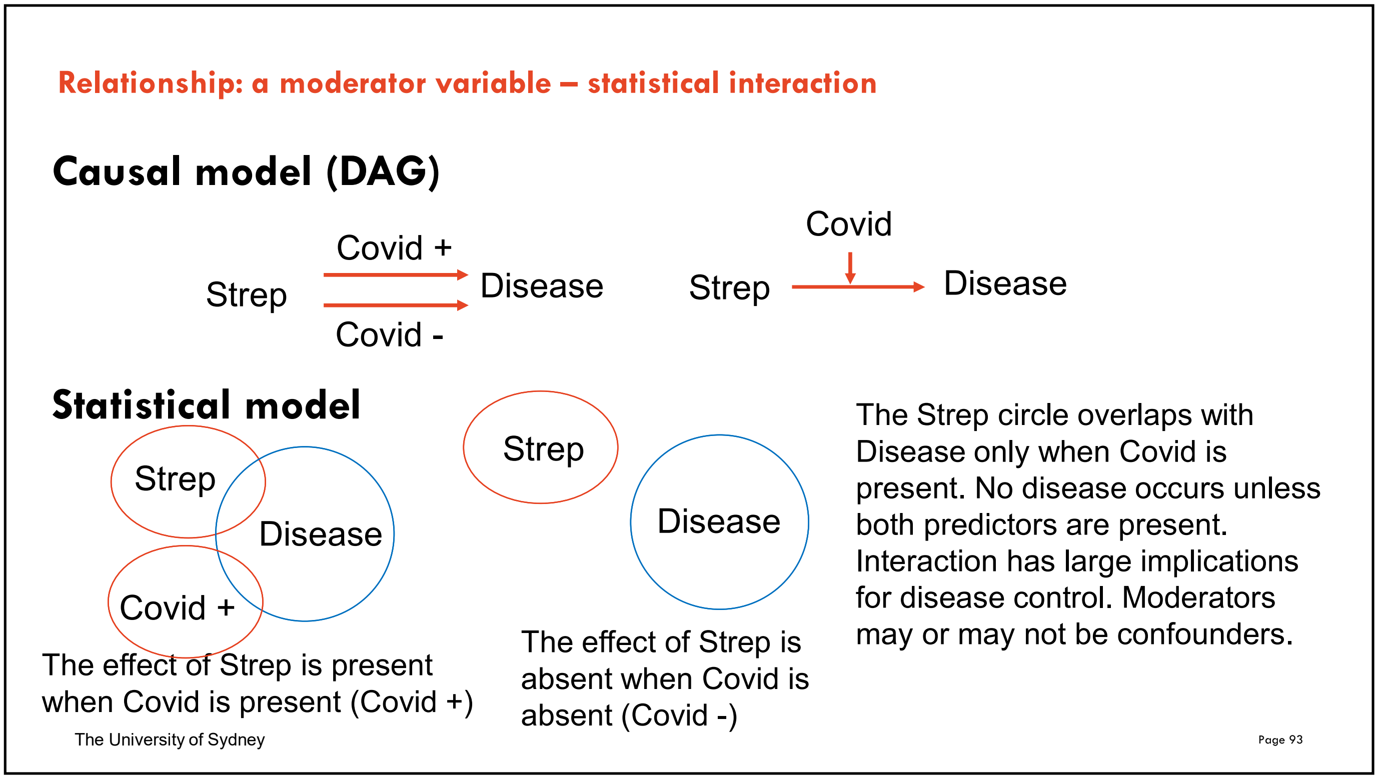
Hospital contact
Strep
Covid
Disease

Strep and the suppressor Covid are both members of the same global variable ‘Hospital contact’ (a proxy variable for exposure to infectious agents)

Statistical model

Disease
Covid
Strep

The variable ‘hospital contact’ has no or only weak unconditional association with the outcome. Once Covid is controlled in the analysis, the Strep circle overlaps with the outcome indicating an association of ‘hospital contact’ with Disease. By controlling the non-causal component of the global variable we reveal and strengthen the suppressed association of the remaining factor with the outcome. The global variable should be refined to exclude Covid. Suppression of the outcome can also occur, e.g. if the disease is not well defined (no association with lung disease but a strong association with a specific type of lung disease).



Relationship: a moderator variable – statistical interaction

Causal model (DAG)

Covid +

Strep → Disease

Covid -

Covid

Strep → Disease

Statistical model

Strep

Disease

Covid +

The effect of Strep is present when Covid is present (Covid +)

Strep

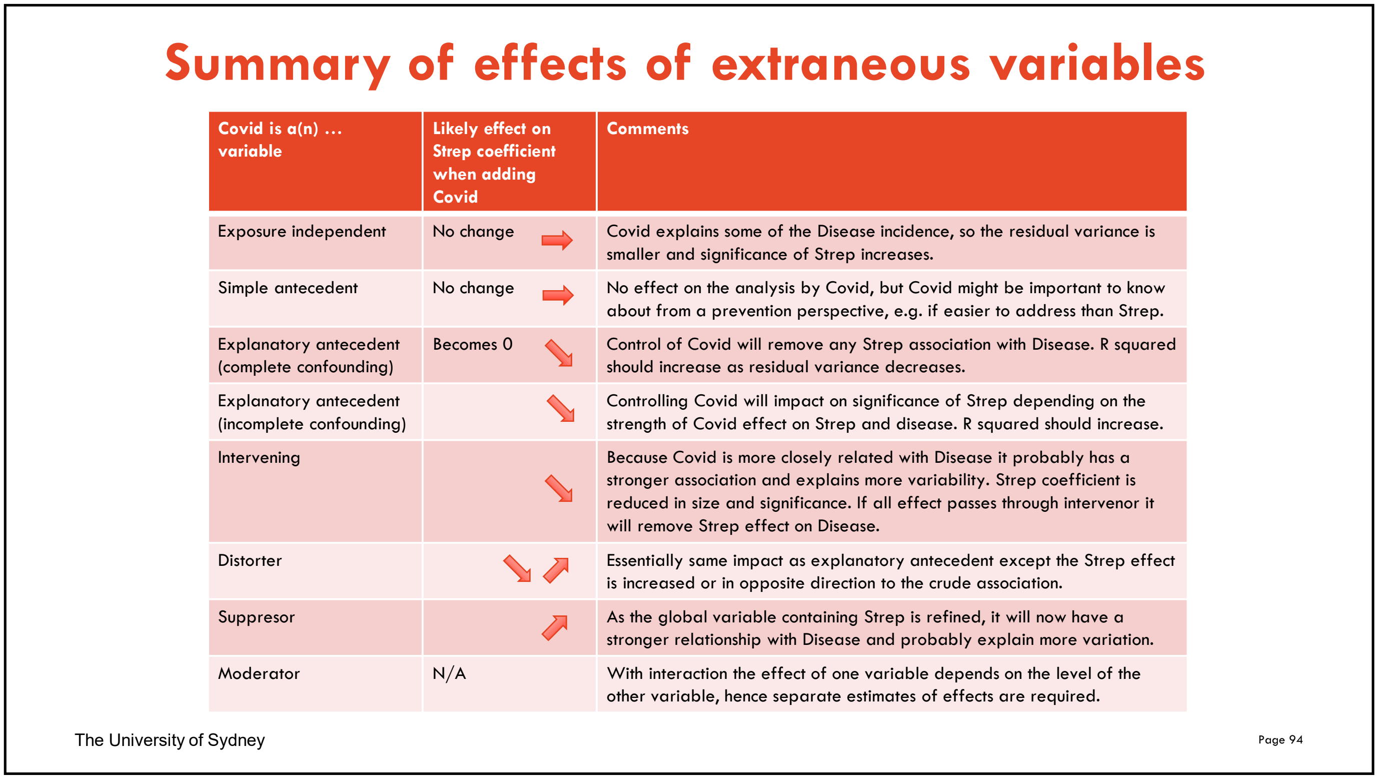
Disease

The effect of Strep is absent when Covid is absent (Covid -)

The Strep circle overlaps with Disease only when Covid is present. No disease occurs unless both predictors are present. Interaction has large implications for disease control. Moderators may or may not be confounders.

The University of Sydney

Page 93



Summary of effects of extraneous variables

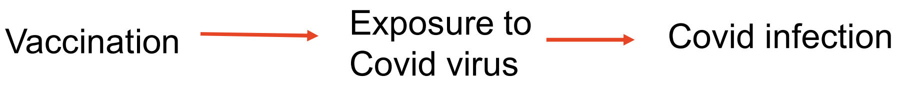
Covid is a(n) … variable	Likely effect on Strep coefficient when adding Covid	Comments
Exposure independent	No change →	Covid explains some of the Disease incidence, so the residual variance is smaller and significance of Strep increases.
Simple antecedent	No change →	No effect on the analysis by Covid, but Covid might be important to know about from a prevention perspective, e.g. if easier to address than Strep.
Explanatory antecedent (complete confounding)	Becomes 0 ↘	Control of Covid will remove any Strep association with Disease. R squared should increase as residual variance decreases.
Explanatory antecedent (incomplete confounding)	↘	Controlling Covid will impact on significance of Strep depending on the strength of Covid effect on Strep and disease. R squared should increase.
Intervening	↘	Because Covid is more closely related with Disease it probably has a stronger association and explains more variability. Strep coefficient is reduced in size and significance. If all effect passes through intervenor it will remove Strep effect on Disease.
Distorter	↘ ↗	Essentially same impact as explanatory antecedent except the Strep effect is increased or in opposite direction to the crude association.
Suppresor	↗	As the global variable containing Strep is refined, it will now have a stronger relationship with Disease and probably explain more variation.
Moderator	N/A	With interaction the effect of one variable depends on the level of the other variable, hence separate estimates of effects are required.

The University of Sydney

Page 94

RCT DAG – what is an ‘instrumental’ variable?

Example: A randomised Controlled Trial (RCT) experiment



Vaccination → Exposure to Covid virus → Covid infection

Treatment:
Control group – not vaccinated
Treatment group – vaccinated

Vaccination is an ‘**Instrumental variable**’, it has direct causal effect on exposure, is unrelated to the outcome and shares no common cause with the outcome