



---

Don't stare into the abyss: understanding your model.

Dr Neil Burns – [neil.burns@sruc.ac.uk](mailto:neil.burns@sruc.ac.uk)

# How the session will work

## Mix of presentation are R coding

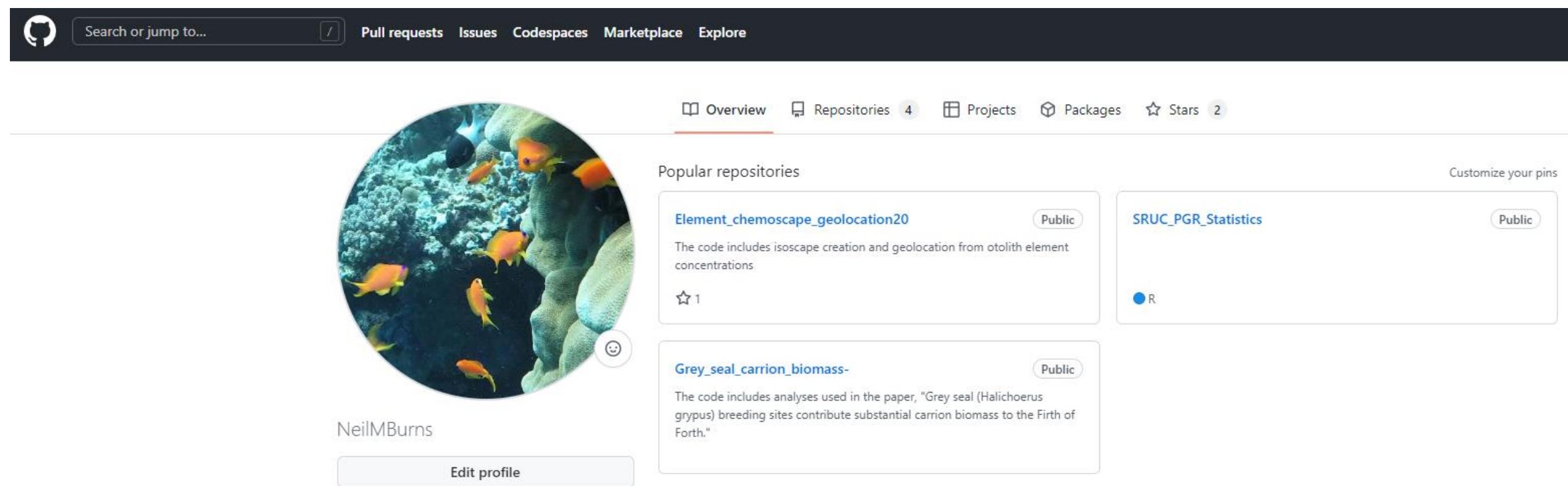
Flit between PowerPoint and R:

- Presentation and code available at my github

<https://github.com/NeilMBurns>

Section headings in R will be highlighted like:

```
#####  
#####      1. some sort of title      #####  
#####
```



The screenshot shows the GitHub profile of NeilMBurns. The profile picture is a circular image of a coral reef with several orange and yellow fish. The username 'NeilMBurns' is displayed below the profile picture, with an 'Edit profile' button underneath. The navigation bar at the top includes 'Pull requests', 'Issues', 'Codespaces', 'Marketplace', and 'Explore'. The main content area shows the 'Overview' tab selected, with a sub-tab for 'Repositories' showing 4 repositories. Below this, there are two 'Popular repositories' listed: 'Element\_chemoscape\_geolocation20' and 'Grey\_seal\_carriion\_biomass-'. Both are marked as 'Public'. To the right, there is a section for 'SRUC\_PGR\_Statistics' also marked as 'Public', with a small R logo below it. A 'Customize your pins' link is visible in the top right corner of the repository section.

Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

Overview Repositories 4 Projects Packages Stars 2

Popular repositories

Customize your pins

**Element\_chemoscape\_geolocation20** Public

The code includes isoscape creation and geolocation from otolith element concentrations

☆ 1

**Grey\_seal\_carriion\_biomass-** Public

The code includes analyses used in the paper, "Grey seal (*Halichoerus grypus*) breeding sites contribute substantial carrion biomass to the Firth of Forth."

**SRUC\_PGR\_Statistics** Public

R

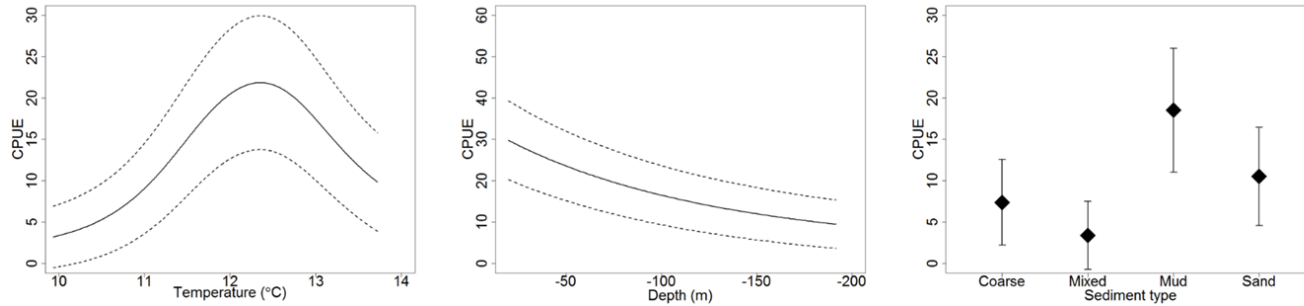
NeilMBurns

Edit profile



# Research Interests

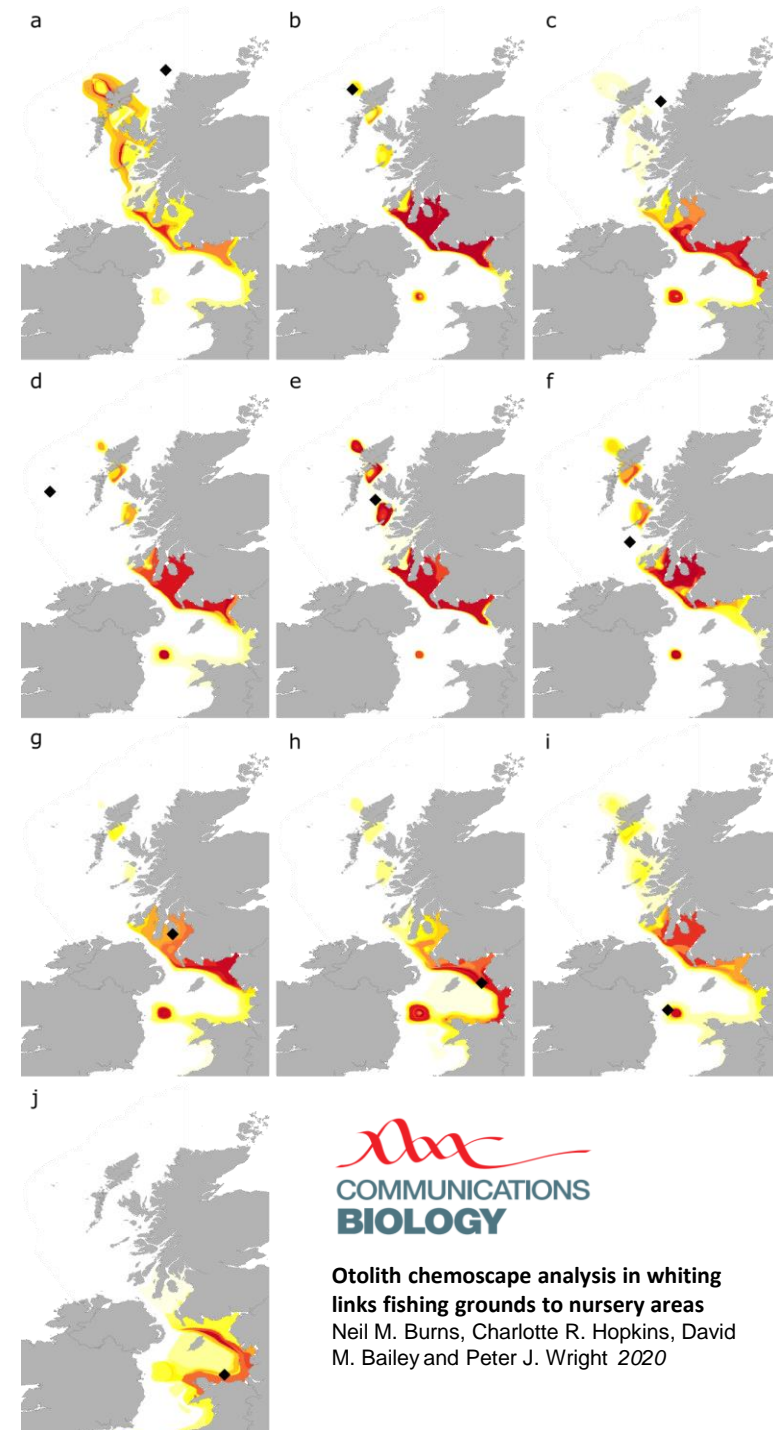
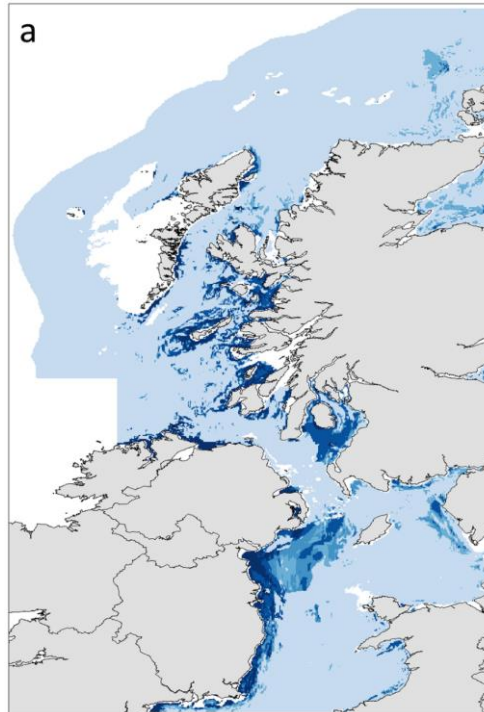
## Population ecology and ecosystem health



RESEARCH ARTICLE

A method to improve fishing selectivity through age targeted fishing using life stage distribution modelling

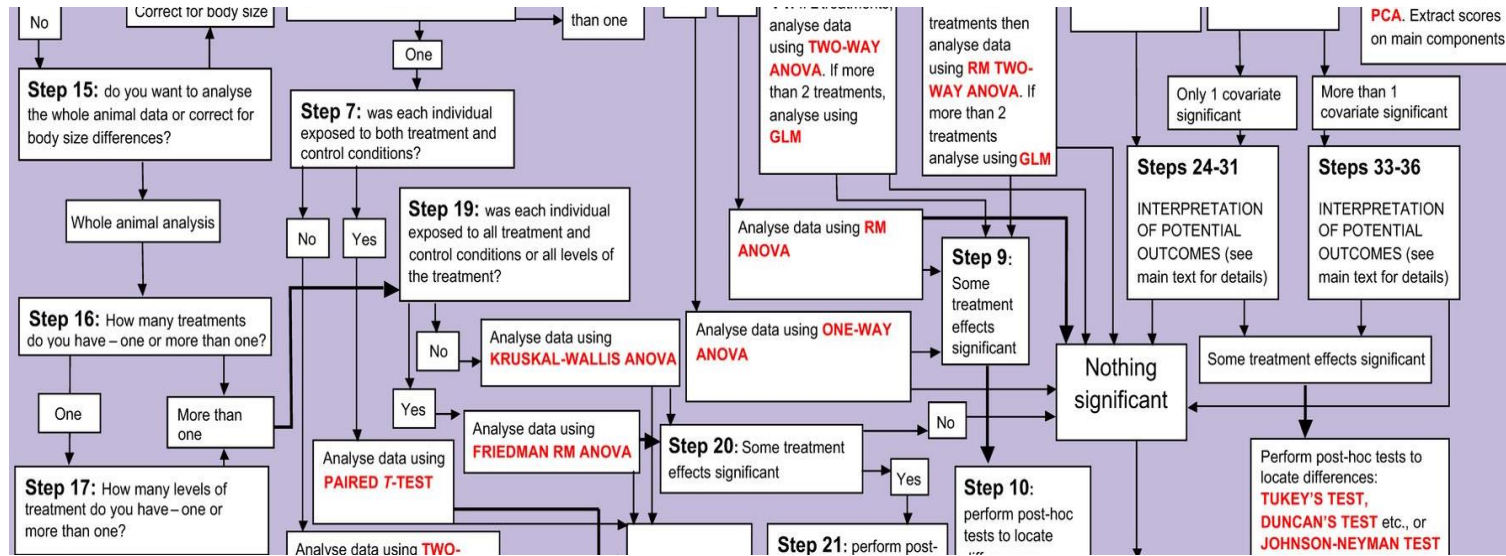
Neil M. Burns<sup>1,2\*</sup>, David M. Bailey<sup>1</sup>, Peter J. Wright<sup>2</sup>



**COMMUNICATIONS BIOLOGY**

Otolith chemoscape analysis in whiting links fishing grounds to nursery areas  
Neil M. Burns, Charlotte R. Hopkins, David M. Bailey and Peter J. Wright 2020

# Why use someone else's robot?



Linear models (lm)  
Generalised Linear Models (glm)  
Generalised Additive Models (gam)  
Mixed effects versions of (lmm, glmm, gamm)

Statistical modelling  
using glms

Why R?



# Building useful robots and making inferences

---

Build the model from the knowledge of your data then use **information theory** to select models (and make inferences).

## A. Model selection

- Picking the best of a bad bunch.

## B. Model validation

- So how rubbish is it?



By using AIC we are selecting the most “cost” effective model. With cost being measured in degrees of freedom.

AIC balance between over and underfitting by estimating out-of-sample deviance without needing to do cross-validation (the gold standard)



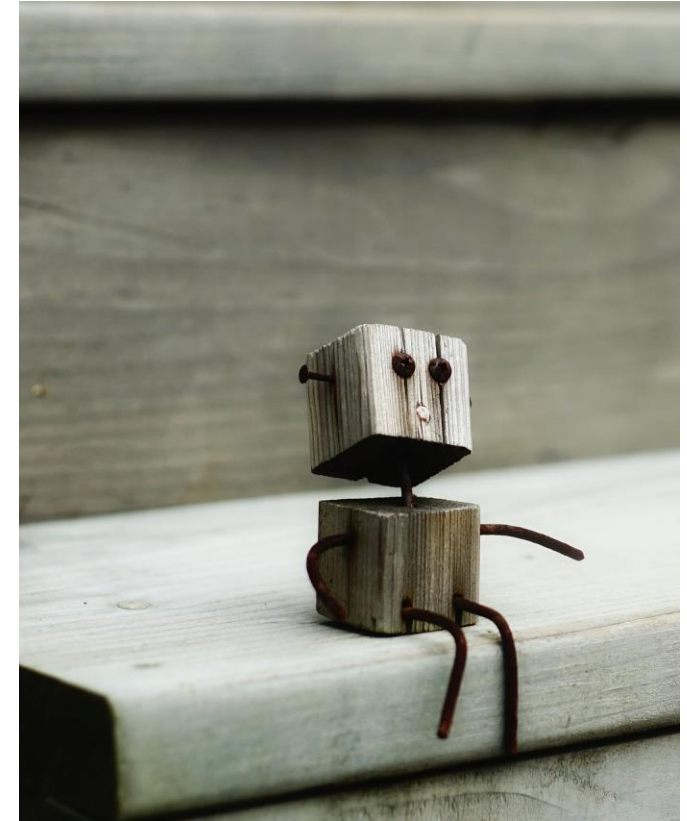
# A - Model selection

---

Using AIC (log-likelihood ratio tests are also useful for those who like a p-value)

## Selection recipe

- Start with the most complex model and work “back” towards the most simple
- Use AIC to choose (3 rules)
  1. Simple models are best
  2. Small AIC is best
  3. If these rules contradict (ie the more complex model has smaller AIC) then AIC should be different by more than 2

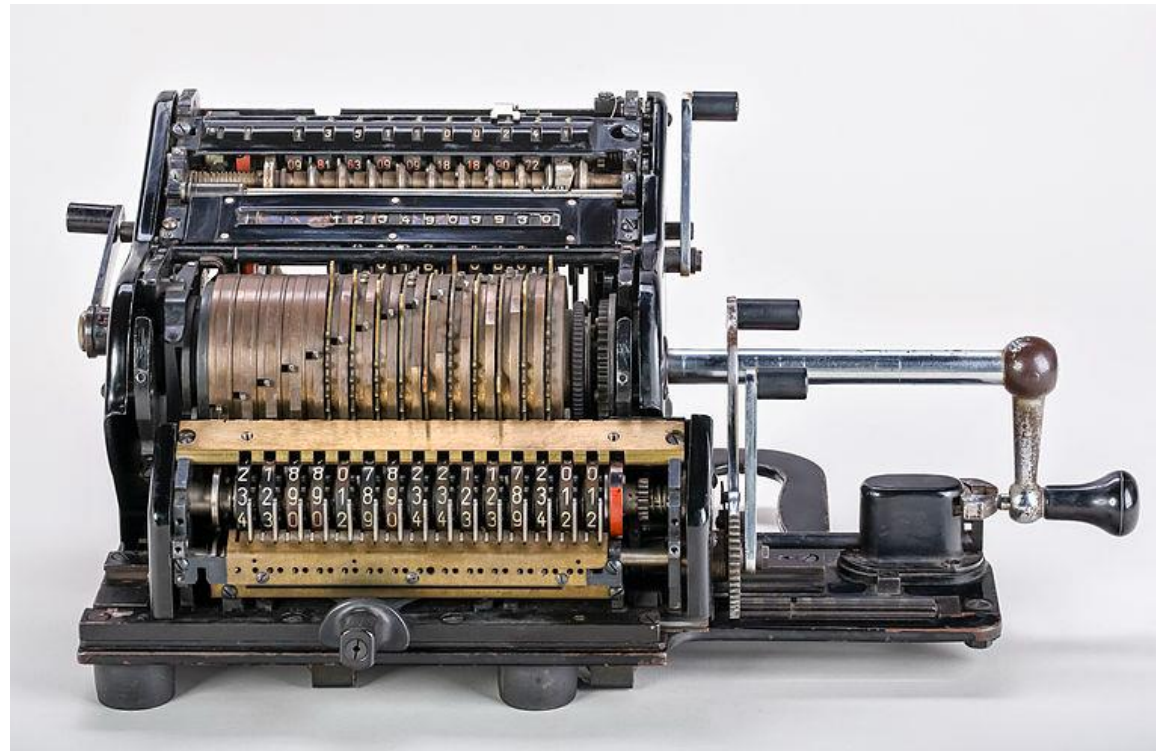


“I used backwards stepwise model selection to ...”

# Final interpretation – a word about staring into the abyss

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.929914   0.417166   4.626 1.14e-05 ***
perc_cov     0.116220   0.007622  15.248 < 2e-16 ***
---
```

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.49784    0.60079   9.151 1.17e-14 ***
perc_cov        0.10071    0.01005  10.026 < 2e-16 ***
cor_colBrown   -1.61468    1.00986  -1.599 0.11320
cor_colGreen    3.69075    1.38125   2.672 0.00889 **
perc_cov:cor_colBrown  0.01524    0.02008   0.759 0.44973
perc_cov:cor_colGreen -0.08092    0.02581  -3.135 0.00229 **
---
```



= 42

# Example 1 – Sharks and coral



Abundance of sharks ( $y_1$ ) ~

It's a whole number (or maybe not)

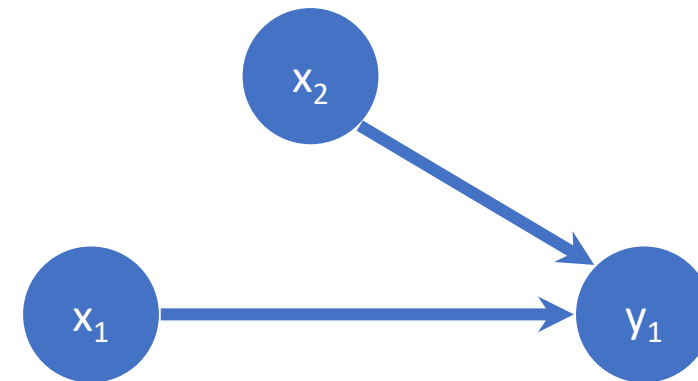
They are big and there are not loads of them so probably less than 60

Coral percentage cover ( $x_1$ )

0 to 100%

Coral colour ( $x_2$ )

Blue, Brown & green



Directed Acyclic Graph (DAG)





Shark  
abundance

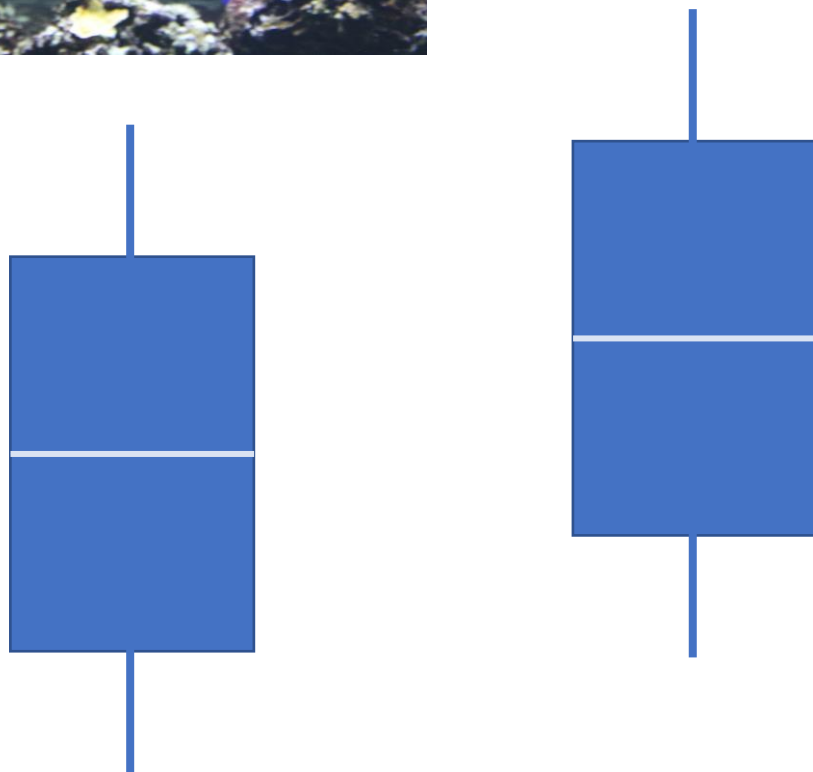


Coral percentage cover





Shark  
abundance



Coral colour



Shark  
abundance

Coral percentage cover





# Back to R – Worked example

```
#####  
#####      2. Understanding our data      #####  
#####
```

```
#####  
#####      3. Models and inference      #####  
#####
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.913513	0.048185	39.71	<2e-16 ***
perc_cov	0.012233	0.000753	16.25	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Coefficients:

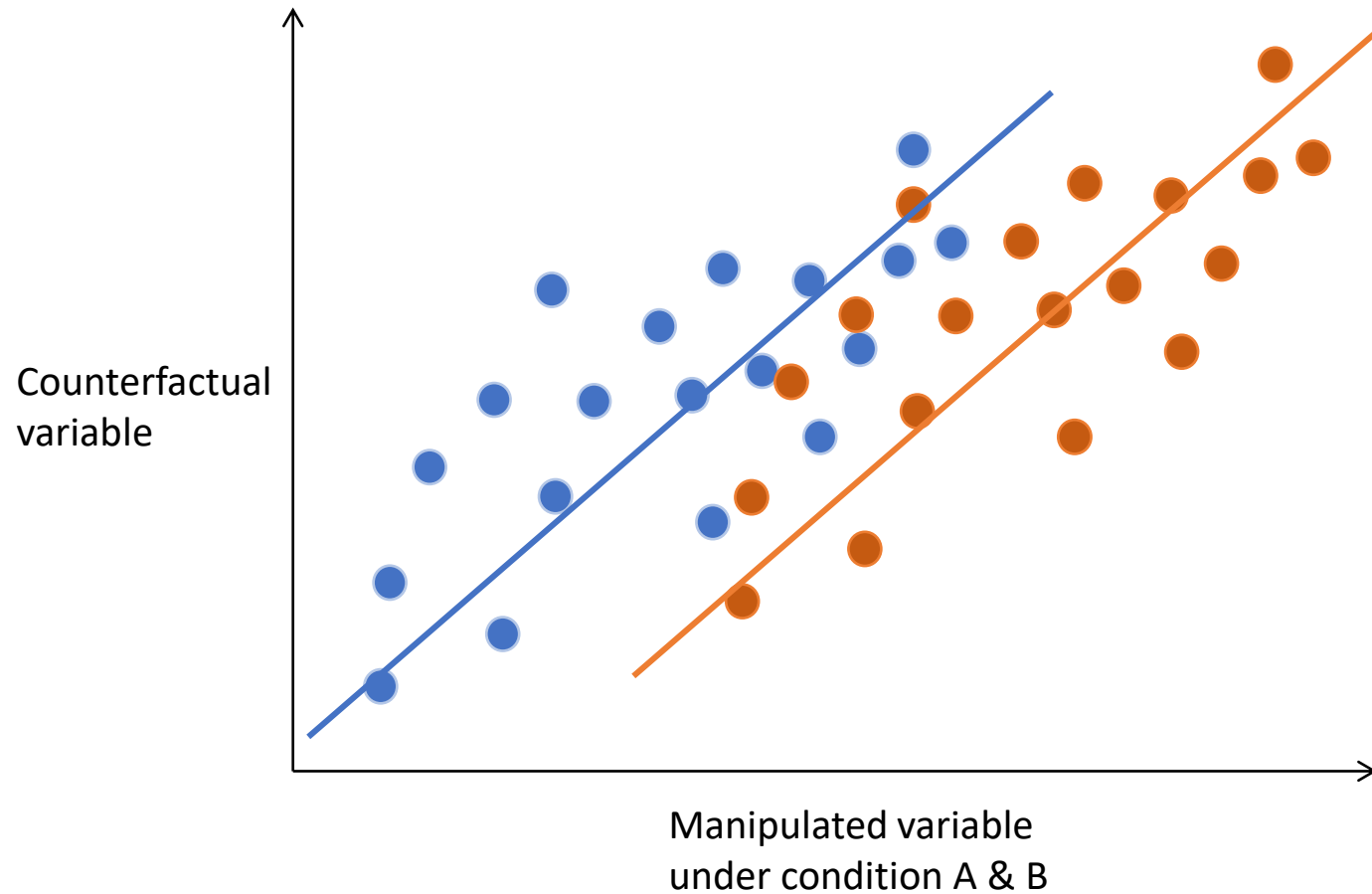
	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	3.5978497	0.0331302	108.597	< 2e-16 ***
perc_cov	0.0196547	0.0004554	43.159	< 2e-16 ***
cor_colBrown	-3.2761745	0.1571768	-20.844	< 2e-16 ***
cor_colGreen	-0.8946106	0.0577938	-15.479	< 2e-16 ***
perc_cov:cor_colBrown	0.0014695	0.0021828	0.673	0.50081
perc_cov:cor_colGreen	0.0027831	0.0008577	3.245	0.00118 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



# Counterfactuals – One continuous variable under a set of conditions (factor/treatment)



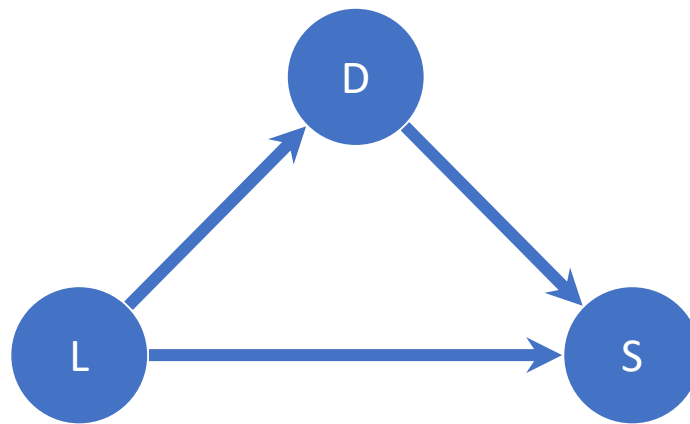
```
#####  
### Use prediction to understand the model ###  
#####
```



## Example 2 – Shark survival



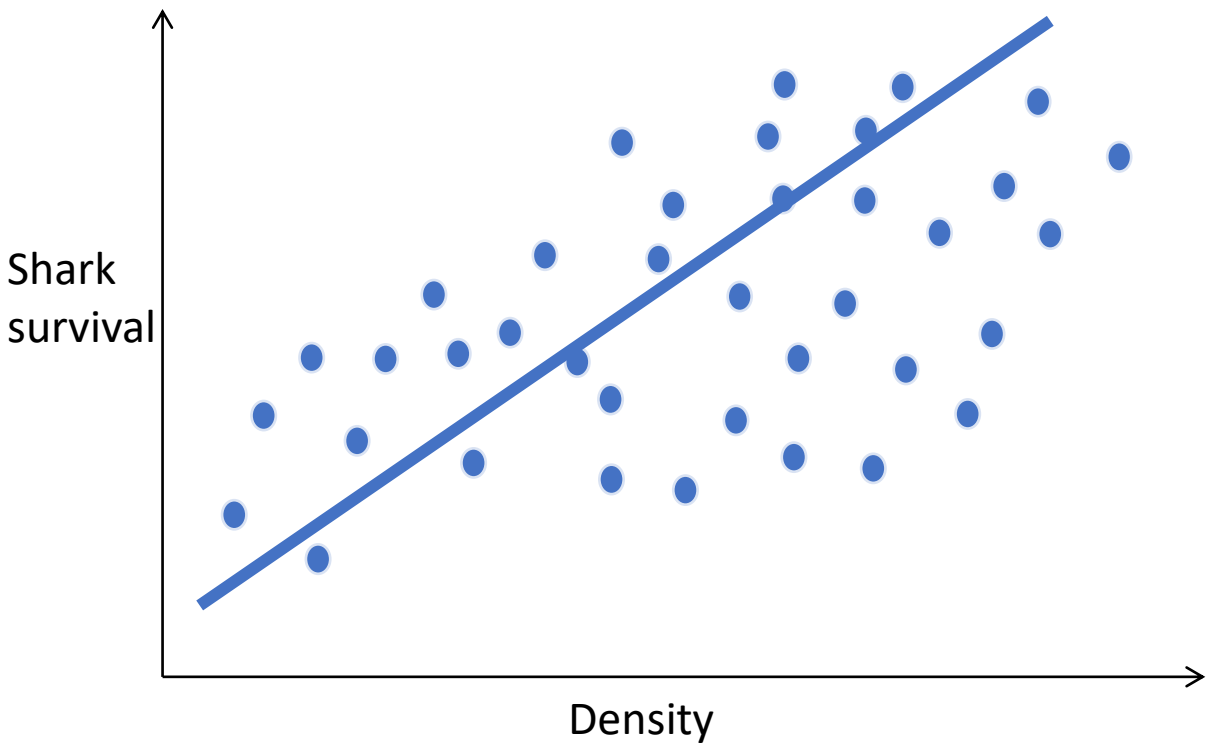
L = shark length  
D = shark density  
S = survival



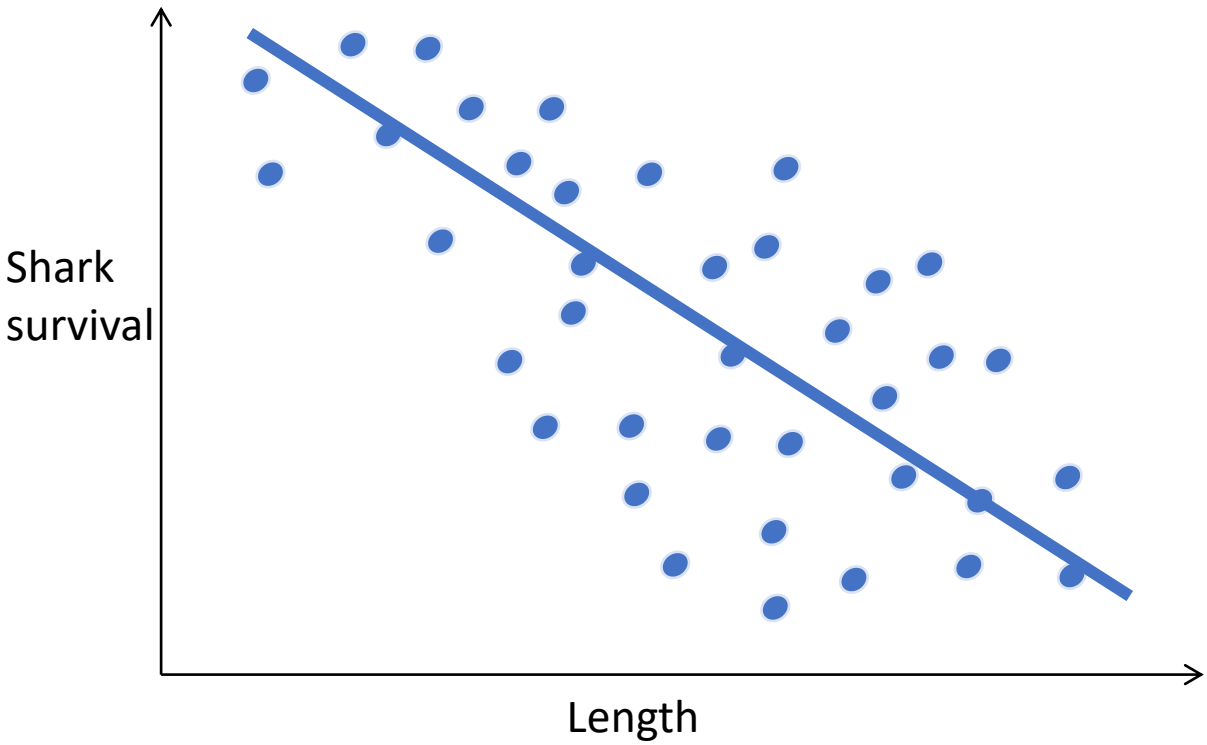


# Method to deal with multiple continuous variable counterfactual

Counterfactual holding length at 3 m



Counterfactual holding density at 5 per km<sup>2</sup>



```
#####  
##### 4. Example 3. Shark survival #####  
#####
```

# Summary

---

- Graph and understand your data (this is critical)
- Most models will be too complex to understand by gazing at coefficient tables
- Use counterfactuals to understand the model (with care)
- Its nice and symmetrical!



# PGR next session?

---

- You direct focus
- Suggestions
  - Continuation of this session
    - GLMs
    - GLMM?
    - GAM?
  - Multivariate modelling
  - Spatial data analysis
  - Writing with impact

