Control, fixed and random variables

Biology **E**xperimental **D**esign and **A**nalysis (**BEDA**)

Januar Harianto

The University of Sydney

Semester 2, 2025

THE UNIVERSITY OF SYDNEY

CRICOS 00026A TEQSA PRV12057

Recall

The general linear model

So far we have been discussing the general linear model (GLM) in terms of coming up with a **relationship** between a **response variable** and one or more **predictors/explanatory variables**.

$$y \sim x_1 + x_2 + \ldots + x_n$$

To simplify today's lecture, we will focus on a simpler model with two predictors (and ***no*** interaction):

$$y \sim x_1 + x_2$$

What model am I fitting?

$$y \sim x_1 + x_2$$

When both predictors are CONTINUOUS

$$\text{body mass} \sim \text{bill length} + \text{flipper length}$$

Use a **regression summary** and interpret beta (β) coefficients:

Characteristic	Beta	95% CI	p-value
bill_length_mm	6.0	-4.1, 16	0.2
flipper_length_mm	48	44, 52	<0.001
Abbreviation: CI = Confidence Interval			

What model am I fitting?

$$y \sim x_1 + x_2$$

When both predictors are CATEGORICAL

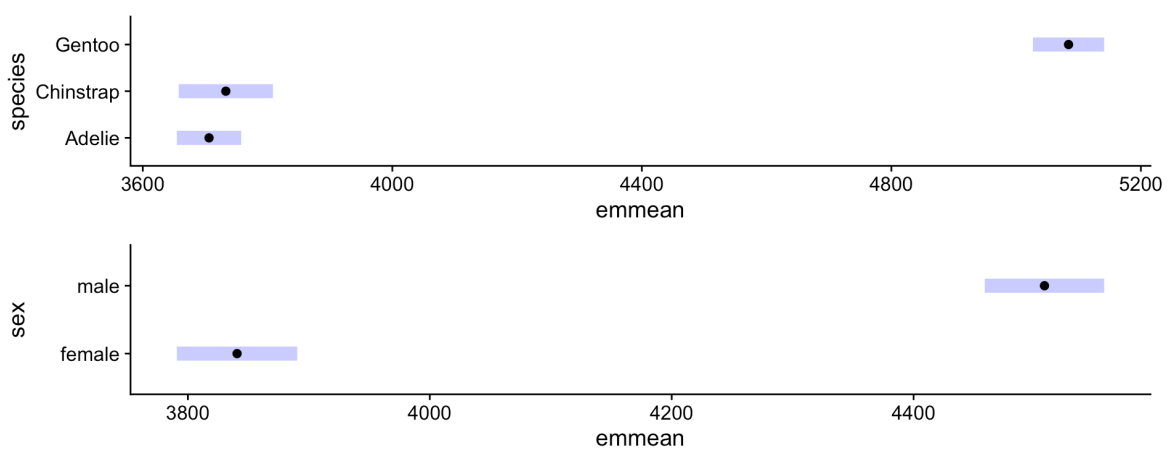
$$\text{body mass} \sim \text{species} + \text{sex}$$

Use an **ANOVA summary** and *also* interpret the **post-hoc tests** to figure out where the differences are:

ANOVA

Effect	DFn	DFd	F	p	p<.05
species	2	329	715.286	1.62e-120	*
sex	1	329	370.012	8.73e-56	*

Estimated marginal means



What model am I fitting?

$$y \sim x_1 + x_2$$

When both predictors are CONTINUOUS and CATEGORICAL

$$\text{body mass} \sim \text{bill length} + \text{species}$$

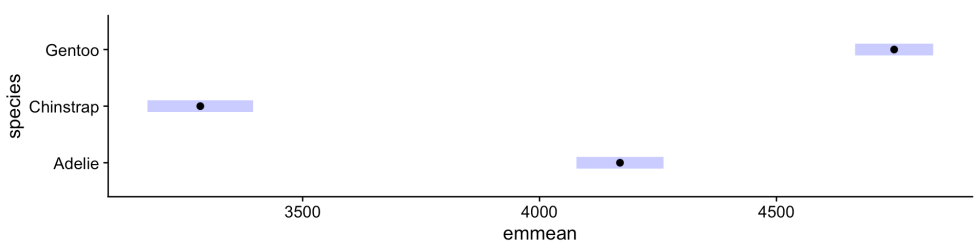
Regression

Characteristic	**Beta**	**95% CI**	**p-value**
bill_length_mm	91	78, 105	<0.001
species			
Adelie	—	—	
Chinstrap	-886	-1,059, -712	<0.001
Gentoo	579	430, 727	<0.001

Abbreviation: CI = Confidence Interval

ANOVA

Effect	DFn	DFd	F	p	p<.05
bill_length_mm	1	338	176.262	1.16e-32	*
species	2	338	333.732	9.71e-81	*



Modelling + study design

> Block what you can; randomise what you cannot.

– George Box (1978), on the design of experiments. Source

The ideal study rarely happens

In the real world, designing a study is *not as simple as fitting a model on just the variables of interest*.

Sometimes…

- Our study is *large and complex* (e.g. spans multiple **locations**, **times**, etc.), so keeping things consistent between them is *almost* impossible.
- We encounter **location-specific** effects that are probably difficult to measure as *specific* variables e.g. farm, lab, greenhouse.
- Other variables can be measured, but are random and not balanced e.g. humidity, sex of the animal sampled, etc.

The ideal study is one where we can isolate these issues in the statistical model so that their effects are "partitioned away" from the variables of interest as *noise*.

The concept of “control” in a model

It turns out that we *can* **plan our data collection** in a way that if certain variables are of not interest but are likely to affect our response variable, we can include their effects in their model as **control variables**.

These variables have special names (but in the end they are just predictors in the model):

- **COVARIATE**: variables that are likely to affect the response variable, yet we **cannot control** them. i.e. they can only be measured. Most likely **continuous**, but can be **categorical**.
- **BLOCK**: variables that are likely to affect the response variable, but we are able to **control** them. i.e. we can manipulate them. Most likely **categorical**.

We include the control variable as a *predictor*, **but we are not interested in its effect on the response variable**.

Covariates

Measured before, during or after the experiment, and not manipulated.

Example – lake species diversity

We want to measure species diversity in lakes (that are influenced by aquaculture) but the size of the lake can affect the diversity of species (e.g. larger lakes can support more species).

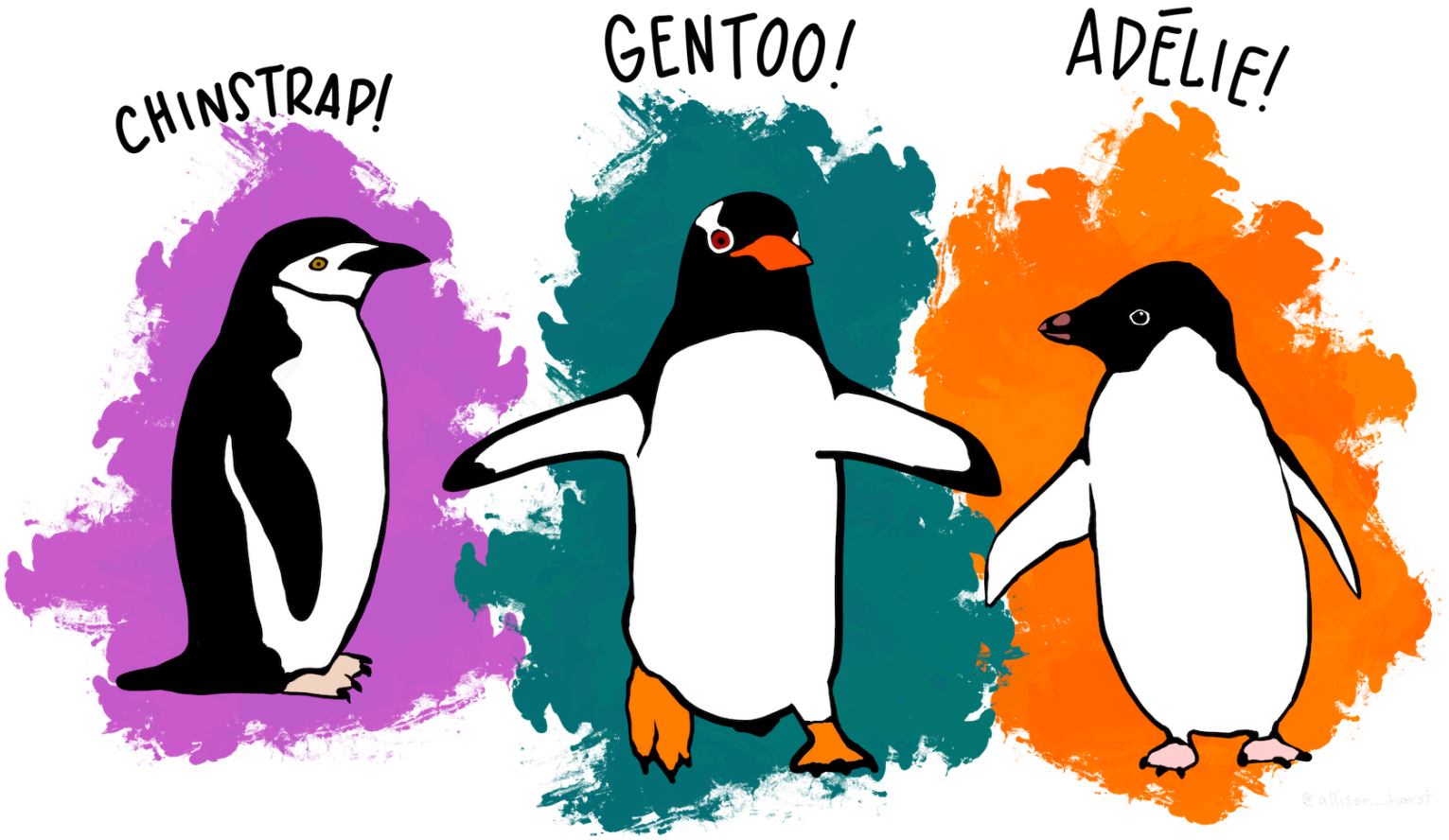
Three Ethiopian lakes. March 29, 2022. Source: NASA Earth Observatory.

$$\text{species diversity} \sim \text{lake size} + \text{species diversity}$$

What is the effect of lake size on species diversity, *if we control for the effects of lake size*?

Example – penguin body mass

We want to determine differences in the body mass of penguins based on their species, but we know that the sex of the penguin matters (e.g. males are generally heavier).



$$\text{body mass} \sim \text{sex} + \text{species}$$

Covariates in the model

$$\text{body mass} \sim \text{sex} + \text{species}$$

In the model, we include the covariate as a **predictor**, but we are not interested in its effect on the response variable.

Characteristic	Beta	95% CI	p-value
species			
Adelie	—	—	
Chinstrap	27	-65, 118	0.6
Gentoo	1,378	1,301, 1,455	<0.001
sex			
female	—	—	
male	668	599, 736	<0.001

Abbreviation: CI = Confidence Interval

We *only* interpret the effect of the species on the body mass. The effect of sex is ignored.

Block

Same concept as covariate, but manipulated by the experimenter, and so is *part* of the study design.

Planning the study to include blocking

Completely randomised design

A	A	C	B
B	A	C	A
B	C	B	C

Distance to shore

Randomised complete block design

A	C	B	**Block 1**
B	A	C	**Block 2**
A	B	C	**Block 3**
C	A	B	**Block 4**

Distance to shore

How blocking works

- We recognise that data may be collected in groups that are location-specific, time-specific, etc.
- “Blocking” on these groups essentially means **repeating the experiment within each group**, and therefore removing within-group variance.
- Thus within-block effects are minimised, allowing us to to focus on between-treatment effects.
- **Doing this requires careful planning of the study design since blocking cannot be implemented/measured after the study is done**.

Example – temporal block

We want to measure the abundance of kangaroos between different habitats, but we know that the time of day can affect the number of kangaroos we see (e.g. more kangaroos are active in the morning). We can use **time of day as a block** and ensure that we sample each habitat at pre-determined times.

$$\text{kangaroo abundance} \sim \text{time of day} + \text{habitat}$$

Example – “worker” block

We want to determine the influence of a new fertiliser on plant growth in a large farm, and so hire multiple workers to apply the fertiliser. We can use the **worker as a block** and ensure that each worker applies the fertiliser to each plot.

$$\text{plant growth} \sim \text{worker} + \text{fertiliser}$$

Example – “spatial” block

We want to measure the effect of different feeds on the growth of pigs in farms. We can use the **farm as a block** and ensure that each feed is given to pigs in each farm.

$$\text{pig growth} \sim \text{farm} + \text{feed}$$

Blocking a model

Just include the block as a predictor in the model, and interpret the treatment effects (here we assume that we intentionally surveyed several islands to consider them as blocks)

$$\text{body mass} \sim \text{island} + \text{species}$$

Characteristic	Beta	95% CI	p-value
species			
Adelie	—	—	
Chinstrap	45	-120, 209	0.6
Gentoo	1,366	1,206, 1,527	<0.001
island			
Biscoe	—	—	
Dream	-21	-205, 162	0.8
Torgersen	-3.3	-191, 184	>0.9

Abbreviation: CI = Confidence Interval

Missing data in blocking

- **Blocking** is a powerful tool in the design of experiments, but it can be difficult to implement in practice.
- Although the idea of blocking is to repeat all treatments within each block, it is possible that some treatments are missing in some blocks, e.g.:
  - A worker is sick and cannot apply the fertiliser.
  - A farm is inaccessible due to flooding.
  - A time of day is missed due to equipment failure.

Problem

- If we have missing data in blocking, that is ok.
- But if we have mising treatment levels in a block we, cannot estimate the variance of the block effect effectively!
- We can still block the data… by defining the block as a **random effect** in the model – i.e. think of it as a **covariate** that we cannot control.

Random effects

Implementation

We assume that the block is a random effect, and that the levels of the block are a random sample from a larger population of possible levels.

$$\text{kangaroo abundance} \sim \text{time of day} + \text{habitat}$$

$$becomes$$

$$\text{kangaroo abundance} \sim \text{error}(\text{time of day}) + \text{habitat}$$

Therefore, we assume that the sample of time values is a **covariate** that comes from a *larger* population of time of days that are normally distributed. In some cases we see the formula written as:

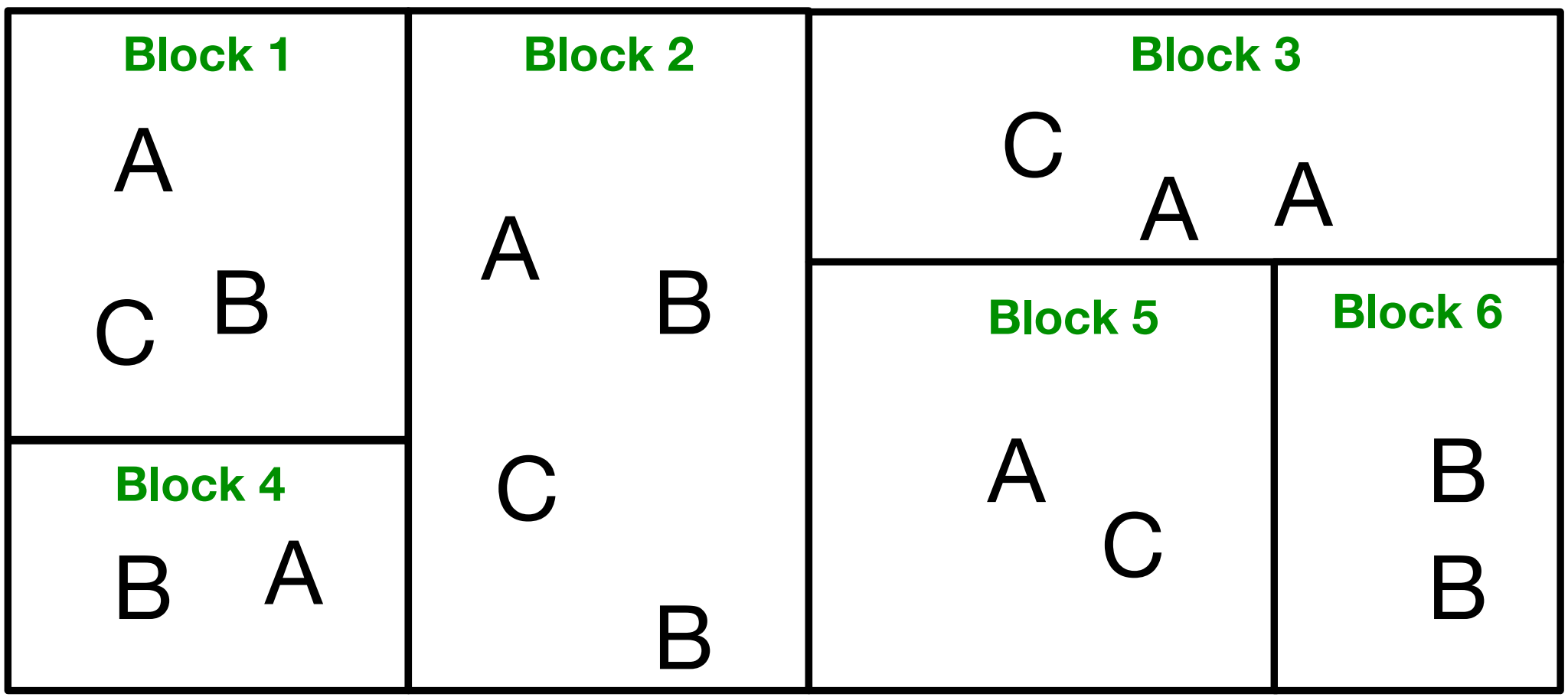
$$\text{response} \sim (1|\text{block}) + \text{treatment}$$

The statistical model will no longer assign a coefficient to the block, but will estimate the variance of the block effect. This variance is then used to estimate the standard error of the treatment effect. **Basically, the block is now *lumped* into the error term.**

$$\text{response} = \beta_0 + b \cdot \text{block} + \beta_1 \, \text{treatment} + \epsilon$$
$$\text{where } b \sim \mathcal{N}(0, \sigma_b^2)$$

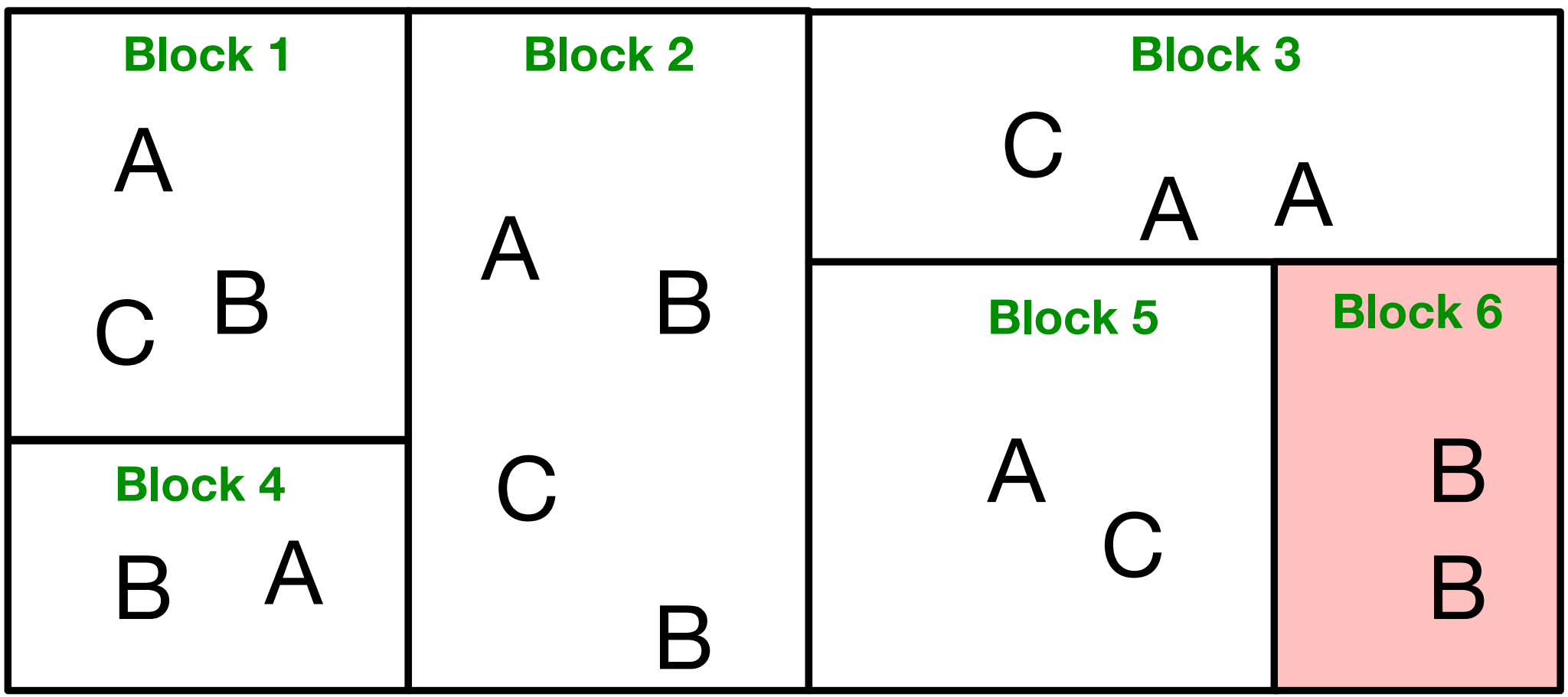
Some variance within blocks is necessary

- For random block effects to work, we need a minimum of **two levels** of treatment within each block.
- If we only have one level of treatment within each block (e.g. all the controls are within a block), we cannot estimate the variance of the block effect, and so the model will not converge!
- One way to fix this is to **combine blocks** to ensure that there are at least two levels of treatment within each block.



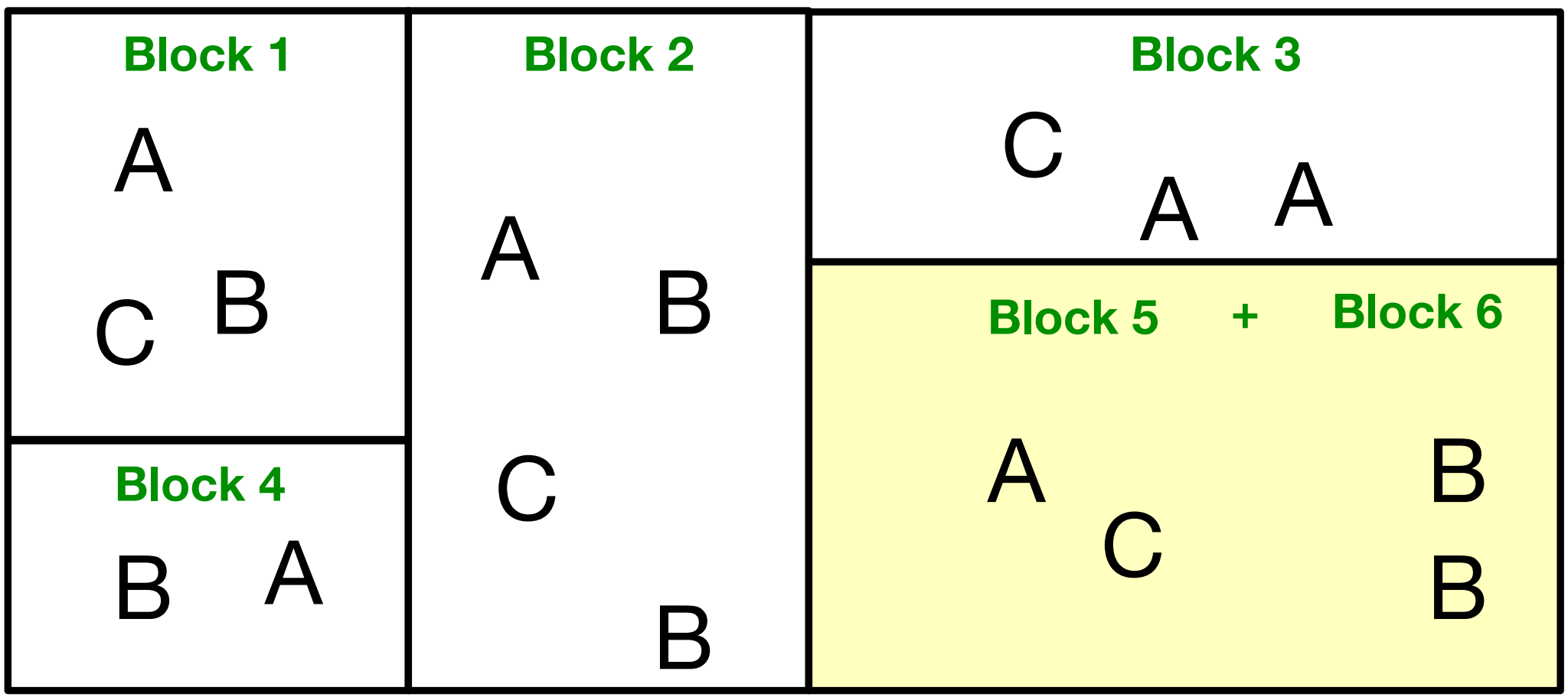
Some variance within blocks is necessary

- For random block effects to work, we need a minimum of **two levels** of treatment within each block.
- If we only have one level of treatment within each block (e.g. all the controls are within a block), we cannot estimate the variance of the block effect, and so the model will not converge!
- One way to fix this is to **combine blocks** to ensure that there are at least two levels of treatment within each block.



Some variance within blocks is necessary

- For random block effects to work, we need a minimum of **two levels** of treatment within each block.
- If we only have one level of treatment within each block (e.g. all the controls are within a block), we cannot estimate the variance of the block effect, and so the model will not converge!
- One way to fix this is to **combine blocks** to ensure that there are at least two levels of treatment within each block.



Let’s wind back…

What model am I fitting?

$$y \sim x_1 + x_2$$

When both predictors are CONTINUOUS, but one is a COVARIATE

$$\text{body mass} \sim \text{error}(\text{flipper length}) + \text{bill length}$$

~~Use a **regression summary** and interpret beta (β) coefficients:~~

Ignore the covariate effect and interpret the remaining coefficients.

Characteristic	Beta	95% CI	p-value
bill_length_mm	6.0	-4.1, 16	0.2
flipper_length_mm	48	44, 52	<0.001
Abbreviation: CI = Confidence Interval			

What model am I fitting?

$$y \sim x_1 + x_2$$

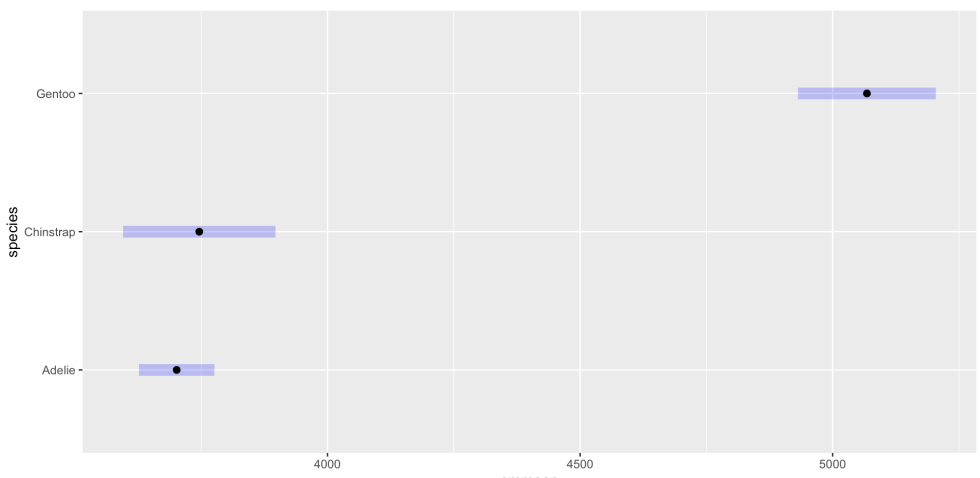
When both predictors are CATEGORICAL, but one is a BLOCK

$$\text{body mass} \sim \text{island} + \text{species}$$

~~Use an **ANOVA summary** and *also* interpret the **post-hoc tests** to figure out where the differences are:~~

Ignore the block effect and interpret the remaining coefficients.

Effect	DFn	DFd	F	p	p<.05
species	2	337	140.894	3.39e-45	*
island	2	337	0.032	9.69e-01	



What model am I fitting?

$$y \sim x_1 + x_2$$

When both predictors are CONTINUOUS and CATEGORICAL but one is a BLOCK

$$\text{body mass} \sim \text{species} + \text{bill length}$$

Characteristic	Beta	95% CI	p-value
bill_length_mm	91	78, 105	<0.001
species			
Adelie	—	—	
Chinstrap	-886	-1,059, -712	<0.001
Gentoo	579	430, 727	<0.001

Abbreviation: CI = Confidence Interval

ANOVA

Effect	DFn	DFd	F	p	p<.05
bill_length_mm	1	338	176.262	1.16e-32	*
species	2	338	333.732	9.71e-81	*

What model am I fitting?

$$y \sim x_1 + x_2$$

When both predictors are CONTINUOUS and CATEGORICAL but one is a COVARIATE

$$\text{body mass} \sim \text{error}(\text{bill length}) + \text{species}$$

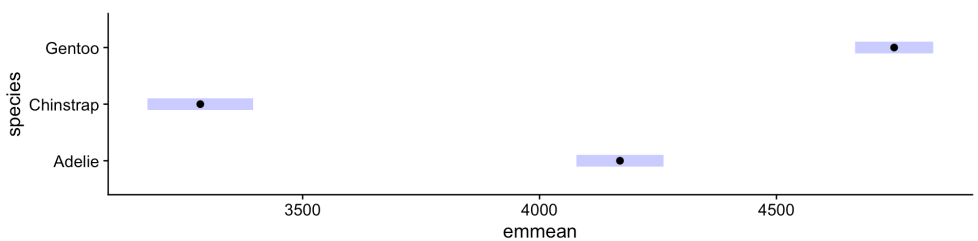
Regression

Characteristic	**Beta**	**95% CI**	**p-value**
bill_length_mm	91	78, 105	<0.001
species			
Adelie	—	—	
Chinstrap	-886	-1,059, -712	<0.001
Gentoo	579	430, 727	<0.001

Abbreviation: CI = Confidence Interval

ANOVA

Effect	DFn	DFd	F	p	p<.05
bill_length_mm	1	338	176.262	1.16e-32	*
species	2	338	333.732	9.71e-81	*



Interpretation – live demonstration in R

… and Jamovi, if there's time

Thanks!

This presentation is based on the SOLES Quarto reveal.js template and is licensed under a Creative Commons Attribution 4.0 International License. A pdf version of this document can be found here.