Modelling multiple differences: a linear model with two or more categorical X variables

BIOL2022 – **B**iology **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

Learning objectives

You should:

- [ ] Frame ANOVA as a specific application of the General Linear Model (GLM) for categorical predictors.
- [ ] Formulate a GLM to test for differences between two or more group means.
- [ ] Explain how an ANOVA summary partitions the variance of a GLM with categorical predictors.
- [ ] Interpret main effects and interaction terms for categorical predictors within a GLM framework.
- [ ] Use post-hoc tests to probe significant main effects and interactions in a GLM.
- [ ] Visualize and interpret the results from a GLM with categorical predictors using interaction plots.

Analysis of variance (ANOVA)

> It's better to solve the right problem approximately than to solve the *wrong* problem exactly.

— John Tukey (1915-2000)

The ANOVA model

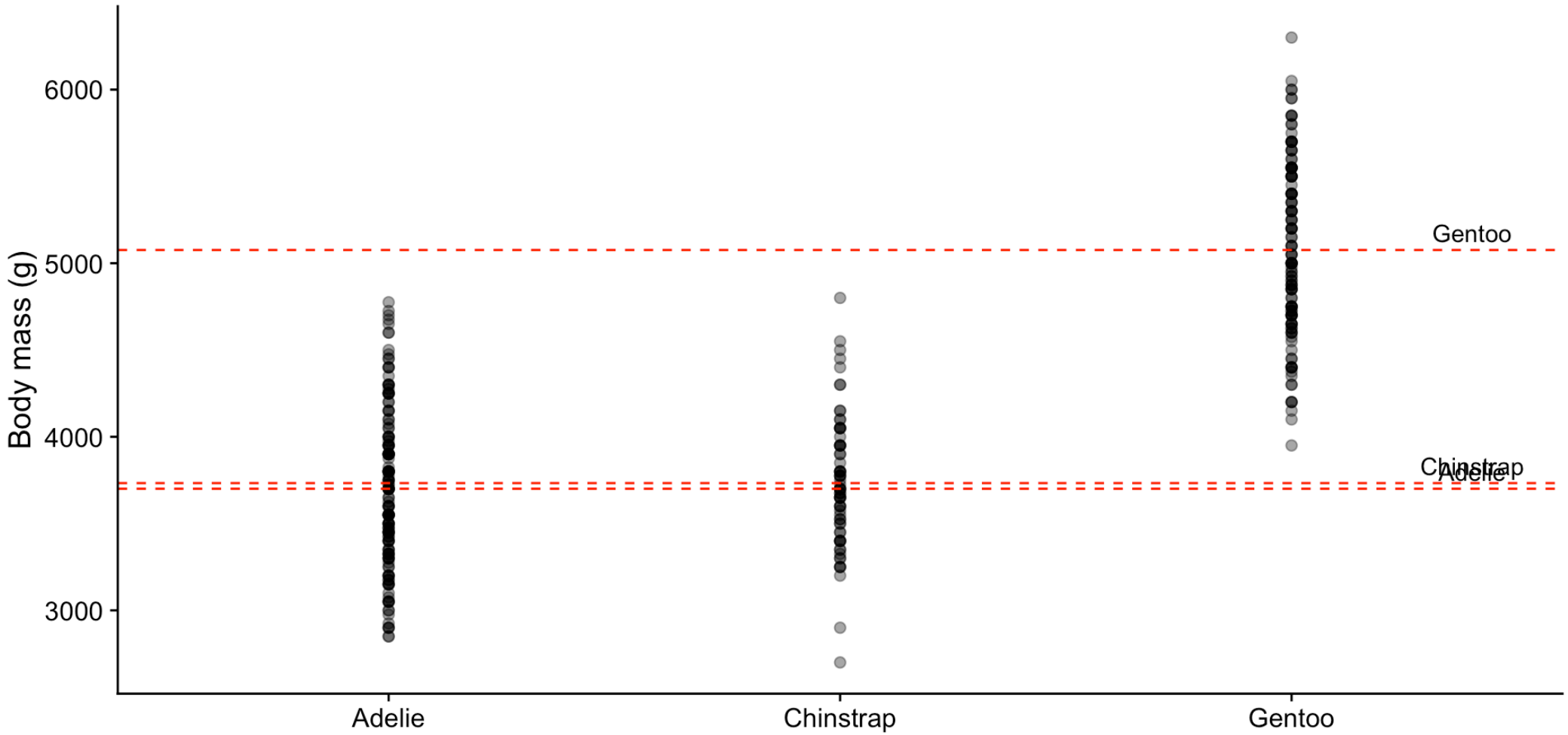
ANOVA is a type of linear regression that uses **categorical** predictors.

How it works (GLM perspective)

1. Fit a linear model using dummy variables for the categories.
2. Split the total variation in the response variable into:
    - Variation *between* groups (explained by the model).
    - Variation *within* groups (unexplained, or residual).
3. Compare these to test if the differences between groups are greater than the variation within groups.

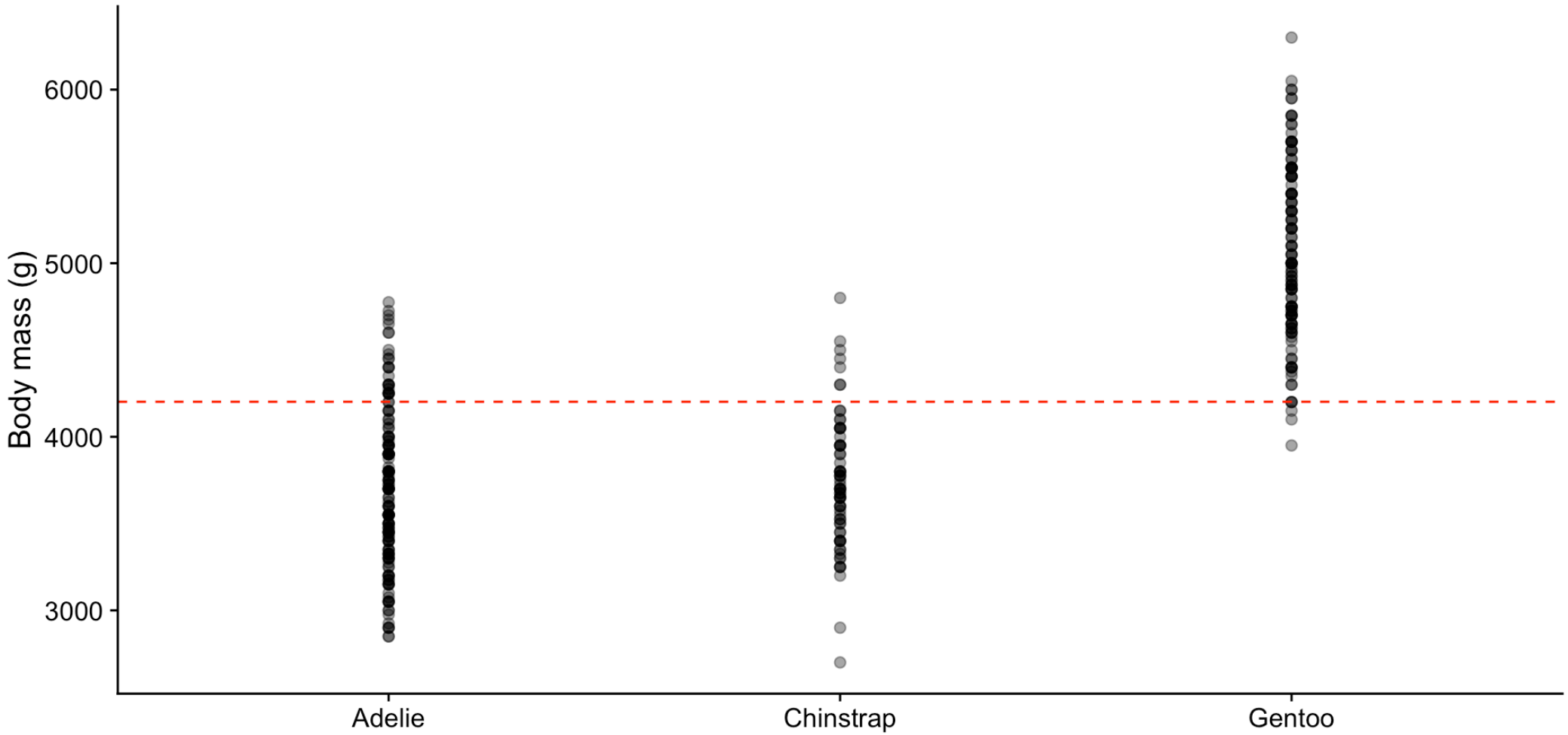
Variation between groups

Can we explain the variation in body mass by species (i.e. **are the means different**)?



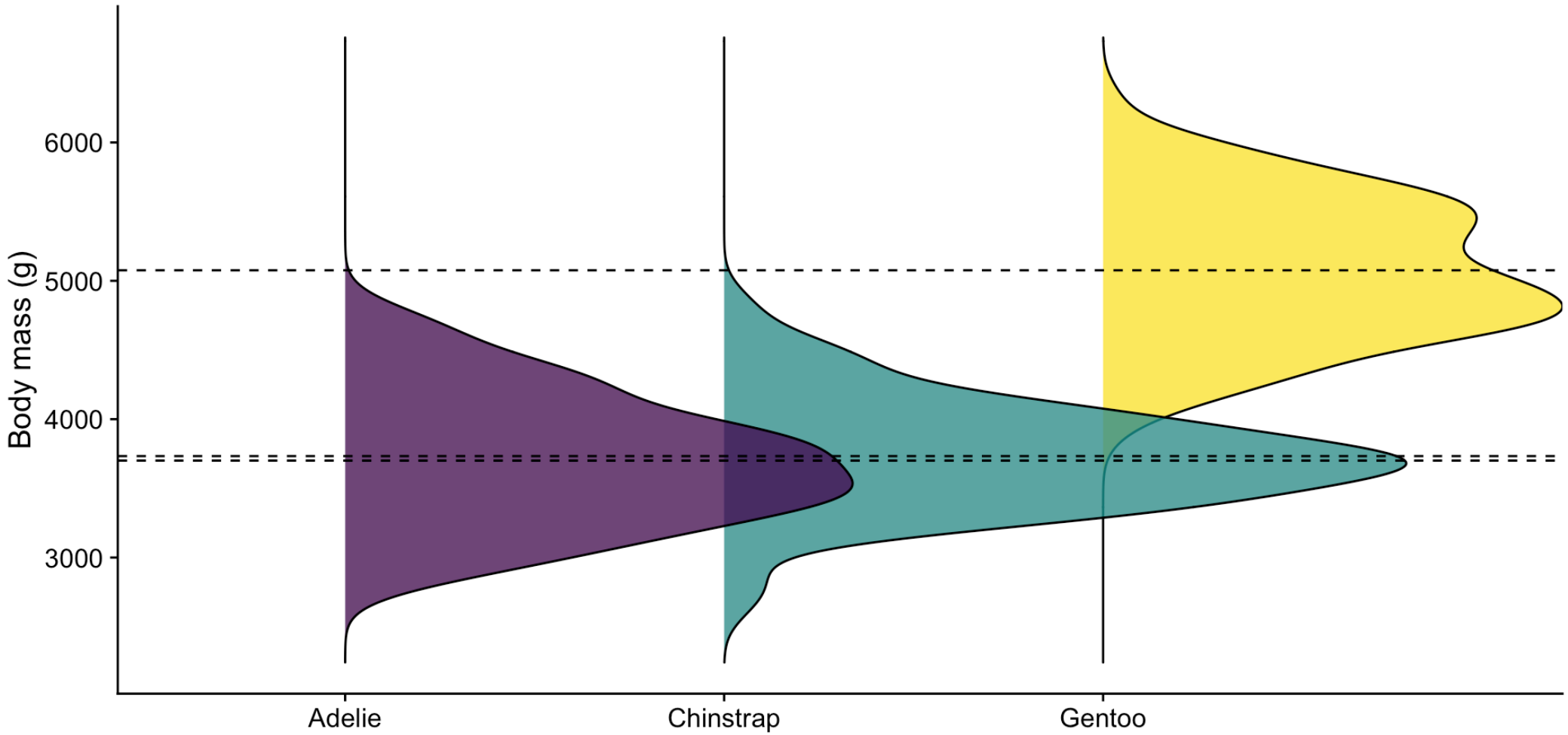
Variation within groups

Can we explain the variation in body mass by an overall mean (i.e. **are they all the same**)?



Another way to look at it

Are the distributions of body mass different enough to be considered separate?



Modelling the relationship

For a GLM:

$$\text{body mass} = \beta_0 + \beta_1 \cdot \text{species} + \epsilon$$

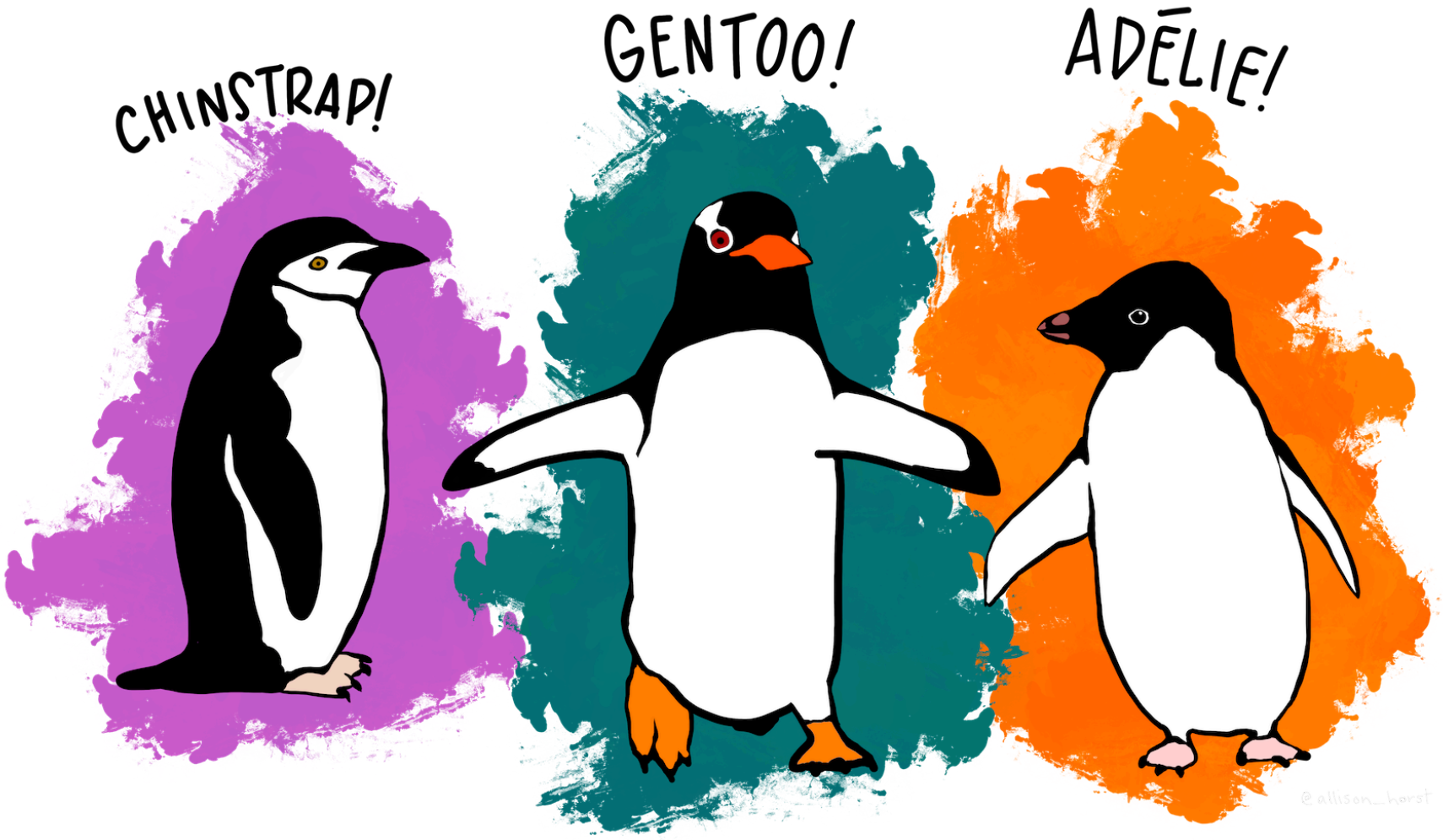
Once we use an ANOVA summary to process the model, we get:

$$\text{body mass} = \text{mean} + \text{species effect (i.e. difference)} + \epsilon$$

That is:

- β_0 is the **overall mean** of body mass.
- β_1 is the **difference** in body mass between species.
- We can add the mean to each species effect to get the **estimated body mass** for each species.

Example



Are species and sex significant predictors of body mass in penguins?

First, the general linear model

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

The interaction term

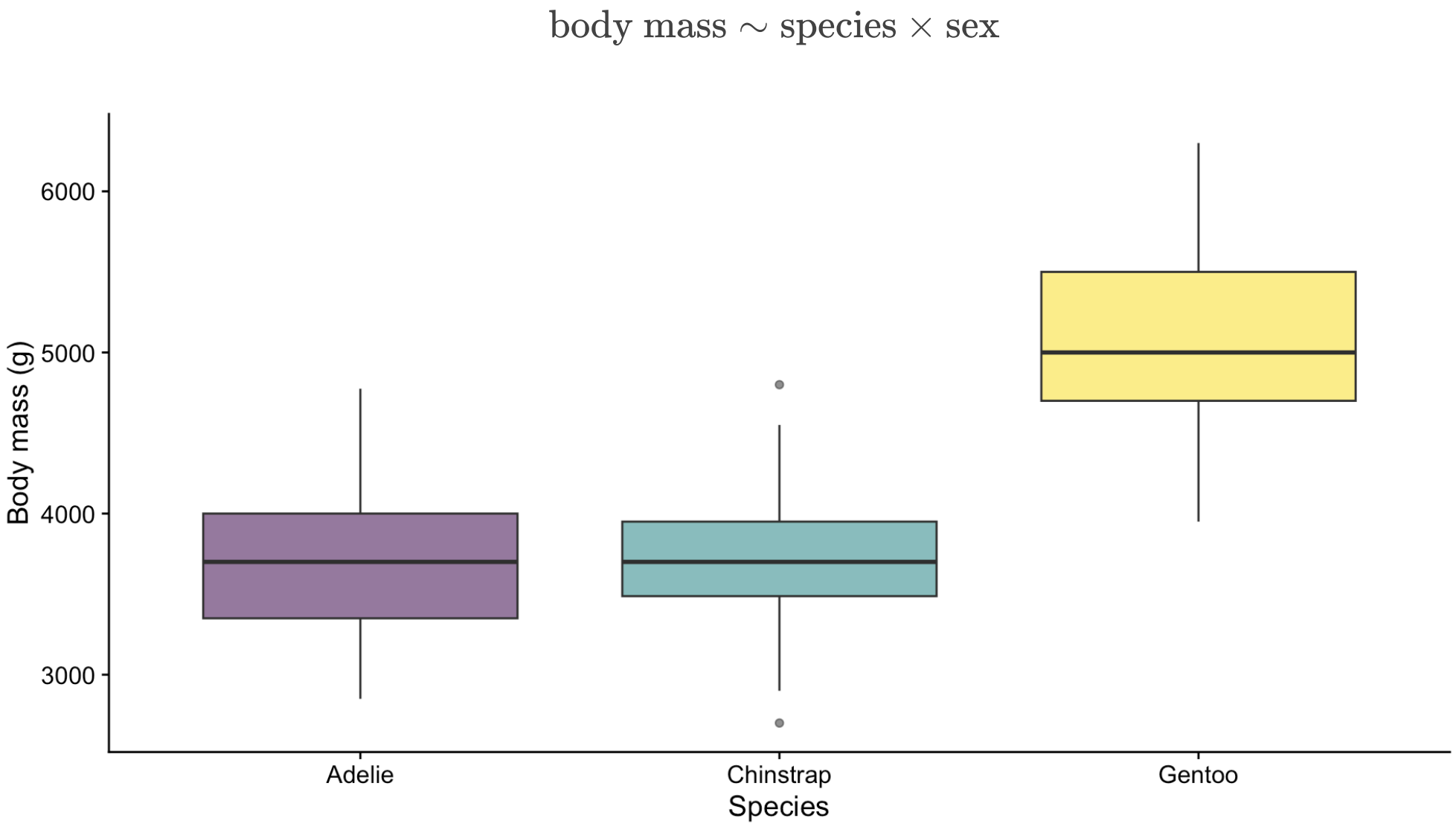
Because we are interested in both predictor variables, we include an **interaction term**:

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

Four steps (as always) - with a post-hoc

1. **Fit the model**, but don't interpret yet. Visualise the relationship (if possible).
2. **Check assumptions** from diagnostic plots (residuals).
3. Select a different model or transform data if assumptions are violated, go back to (2). Skip if assumptions are met.
4. **Interpret** the model ✨**and post-hoc**✨.

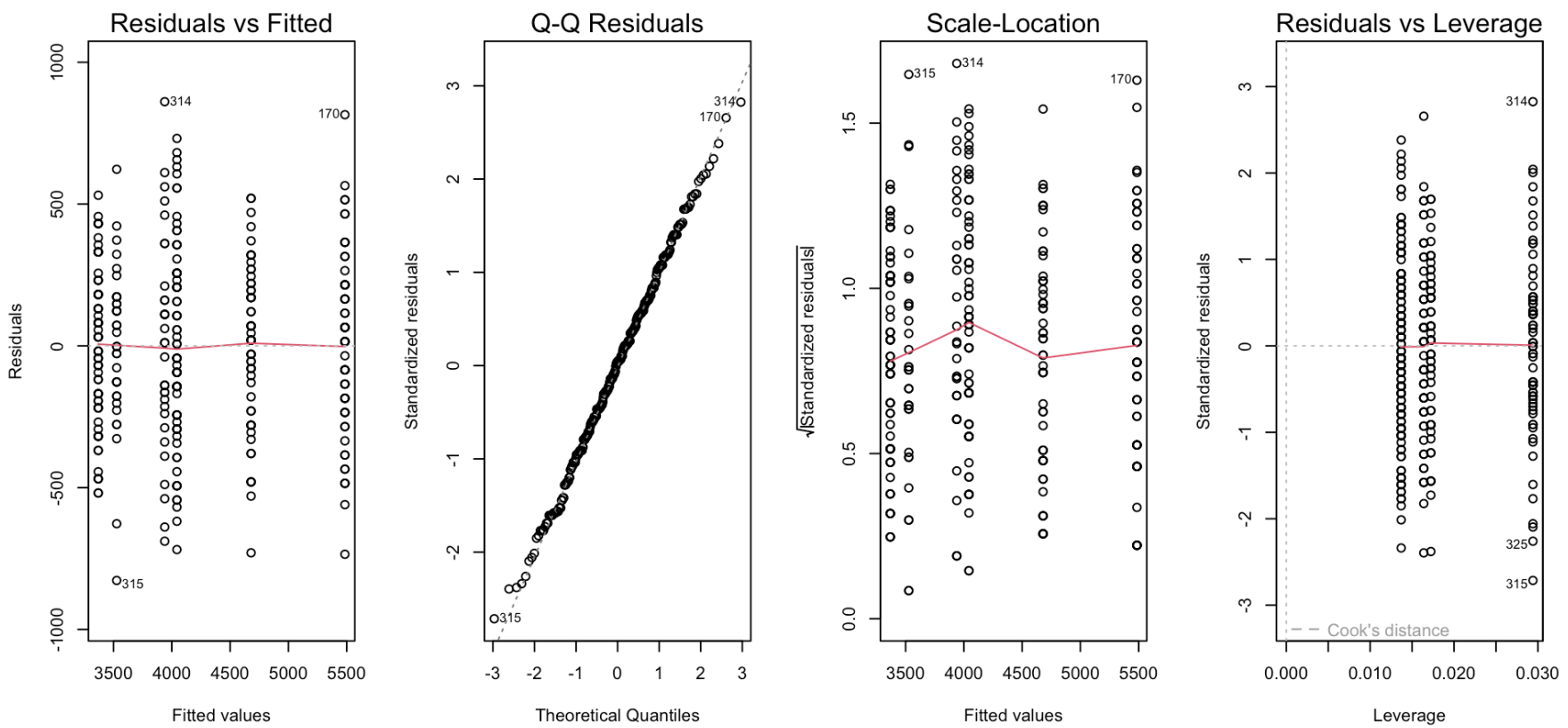
1. Model/plot



Side note: Beyond bar and box plots

2. Check assumptions

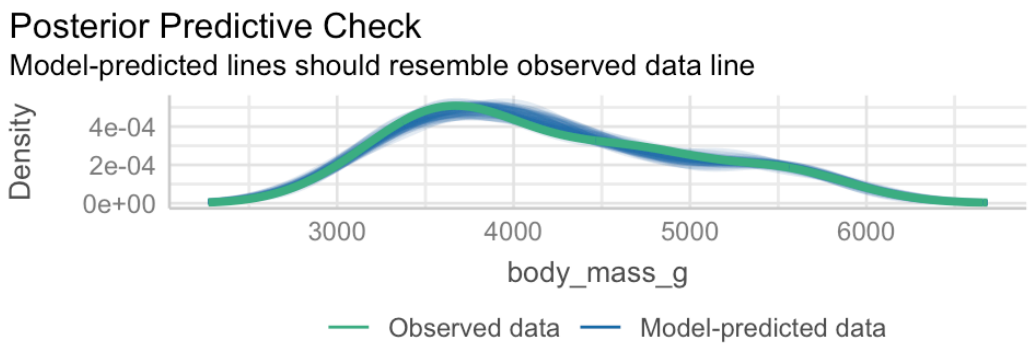
$$\text{body mass} \sim \text{species} \times \text{sex}$$

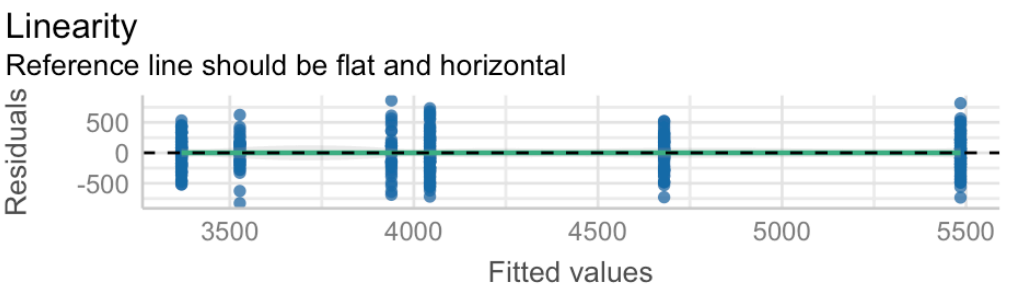


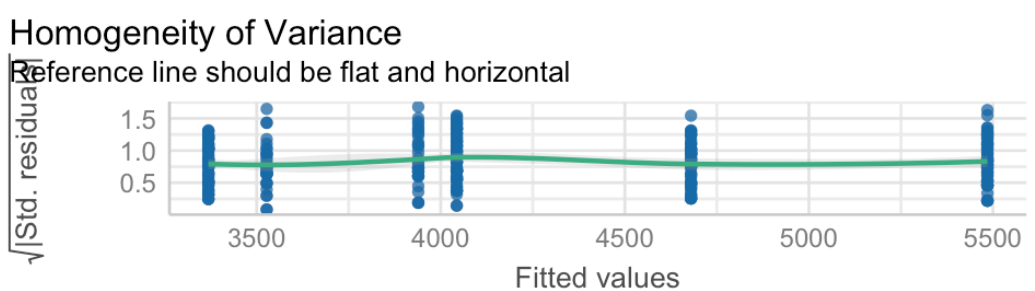
- We see clusters of residuals - a total of 6, representing the 6 combinations of **species** × **sex**.
- Interpretation of the residuals are as normal – just think of **LINE**!

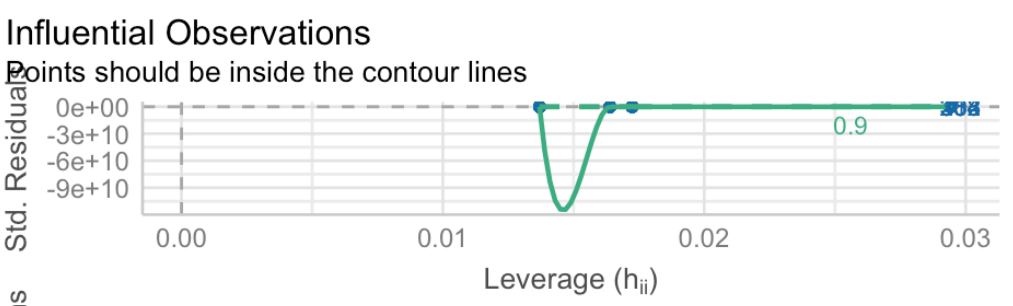
2. Check assumptions

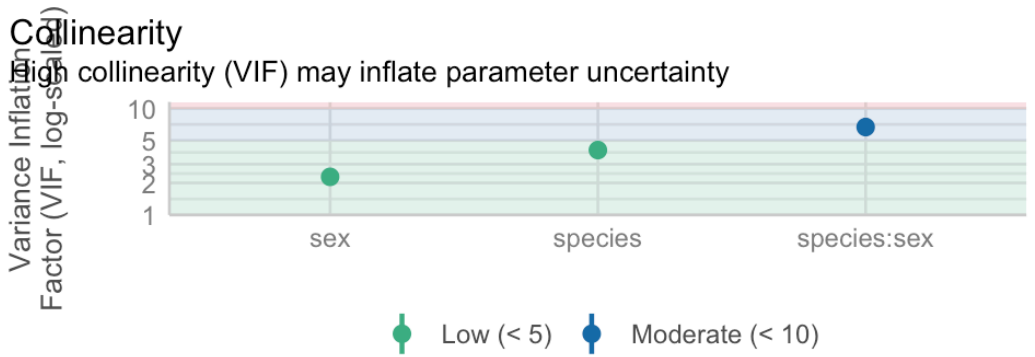
$$\text{body mass} \sim \text{species} \times \text{sex}$$

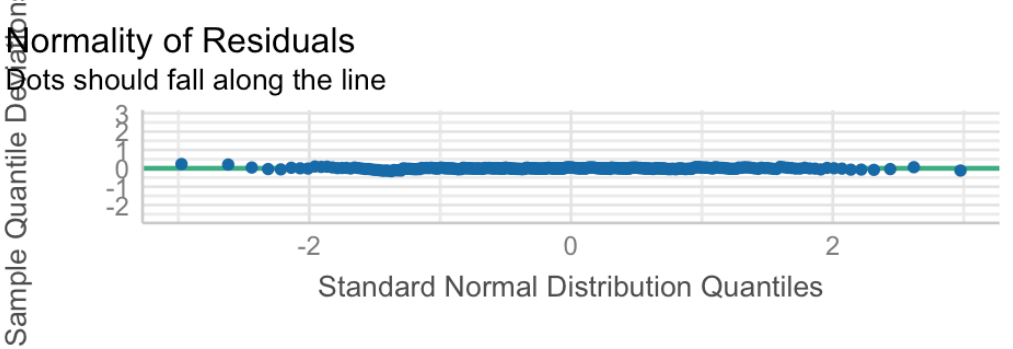












4. Interpret

In practice, if we are interested in comparing means, the **ANOVA summary** table is the most useful.

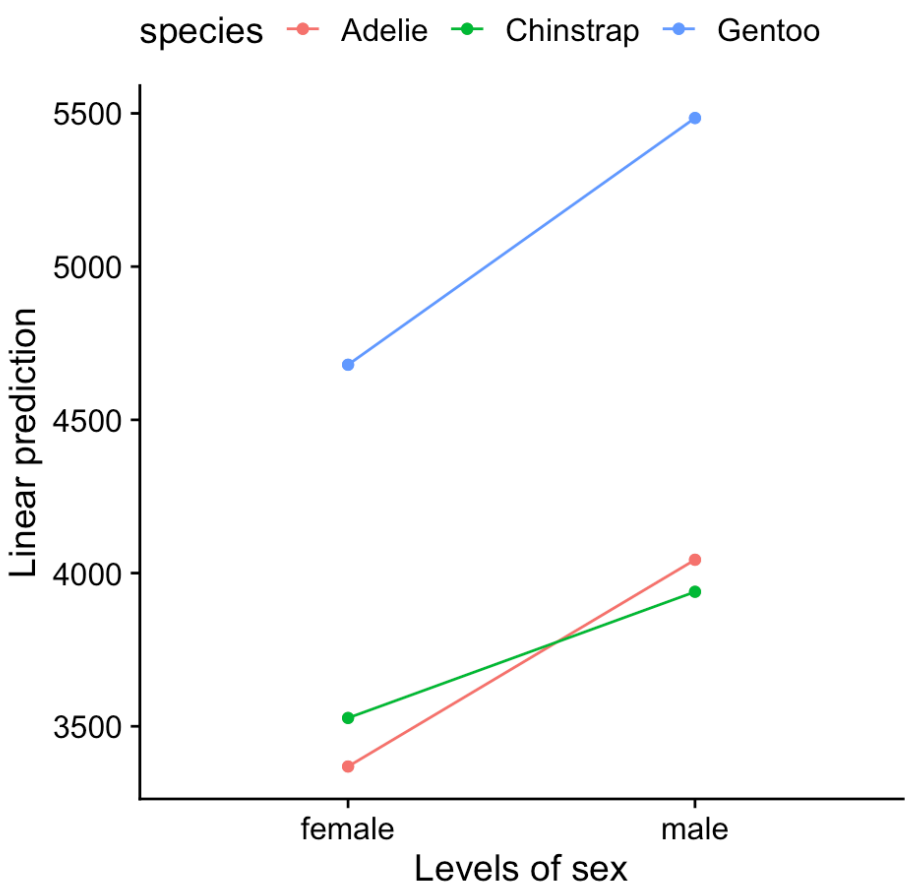
```
Analysis of Variance Table

Response: body_mass_g
             Df    Sum Sq  Mean Sq F value    Pr(>F)
species       2 145190219 72595110 758.358 < 2.2e-16 ***
sex           1  37090262 37090262 387.460 < 2.2e-16 ***
species:sex   2   1676557   838278   8.757 0.0001973 ***
Residuals   327  31302628    95727
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

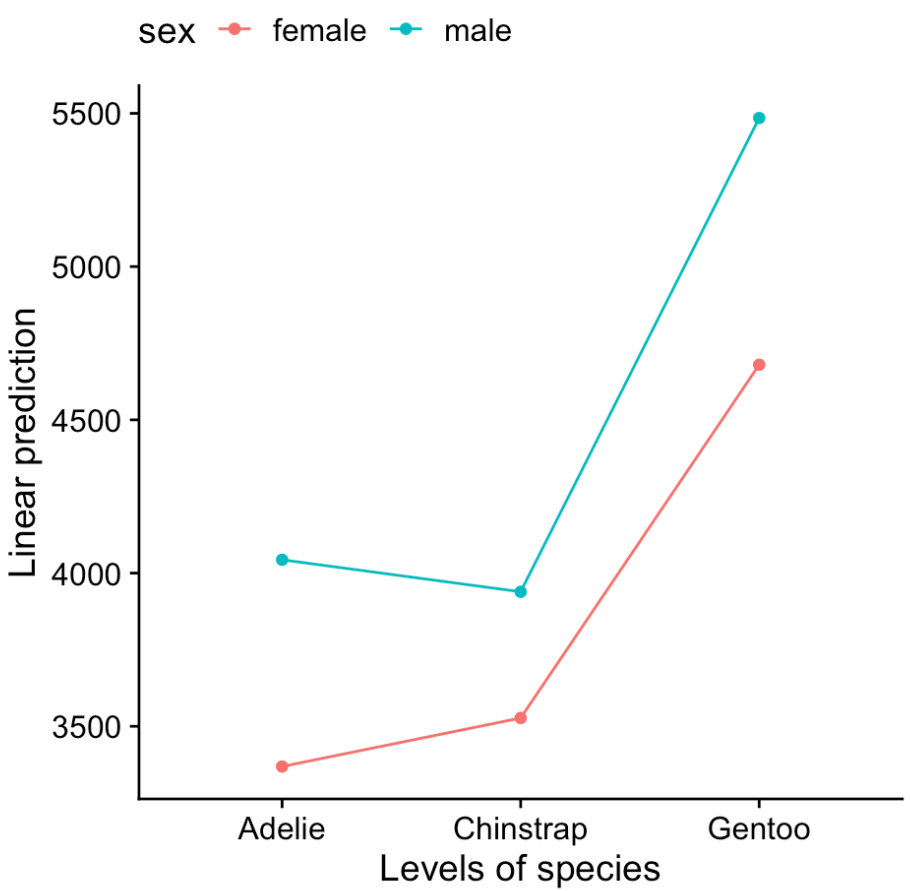
There are **significant interactions**, so we no longer interpret the main effects. **Interpret the interaction term instead.**

Interpreting interactions

We want to check two things:

- Whether the relationship between body mass and species is dependent on **sex**, and/or
- Whether the relationship between body mass and sex is dependent on **species**.





Interpreting interactions

- Species affects the relationship between body mass and sex. Males are generally heavier than females, but the difference is less pronounced in Chinstrap penguins as seen by the flatter slope compared to the other two species.
- Sex also affects the relationship between body mass and species. Adelie males are generally heavier than Chinstrap males, but in females it is the opposite – Adelie females are generally lighter than Chinstrap females.

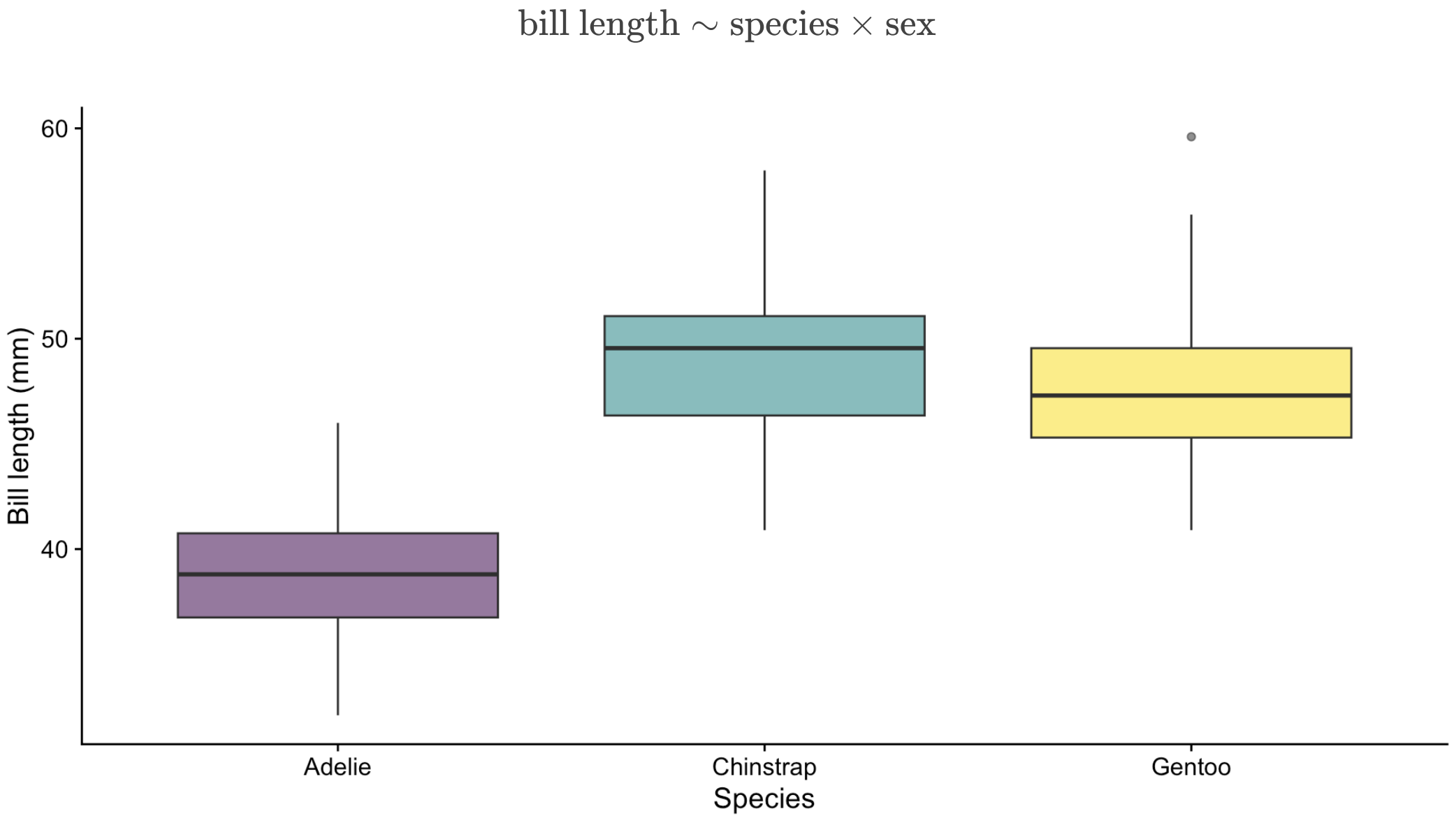
What if we look at the non-ANOVA summary?

```
Call:
lm(formula = body_mass_g ~ species * sex, data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-827.21 -213.97   11.03  206.51  861.03

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             3368.84      36.21  93.030  < 2e-16 ***
speciesChinstrap         158.37      64.24   2.465  0.01420 *
speciesGentoo           1310.91      54.42  24.088  < 2e-16 ***
sexmale                  674.66      51.21  13.174  < 2e-16 ***
speciesChinstrap:sexmale -262.89      90.85  -2.894  0.00406 **
speciesGentoo:sexmale     130.44      76.44   1.706  0.08886 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 309.4 on 327 degrees of freedom
```

Example 2

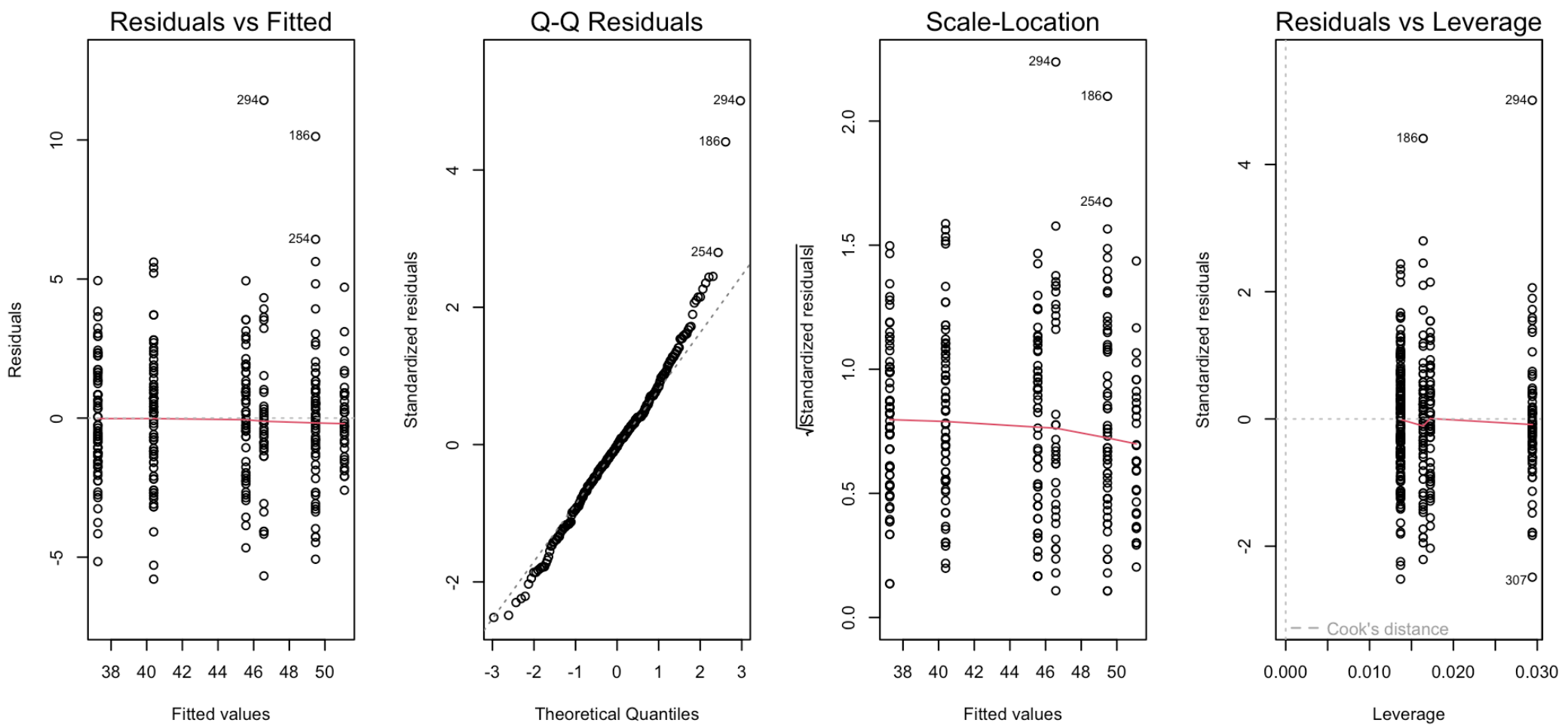
Are species and sex significant predictors of bill length in penguins?

1. Model/plot


bill length ∼ species × sex
60
50
40
Bill length (mm)
Adelie
Chinstrap
Gentoo
Species

2. Check assumptions

$$\text{bill length} \sim \text{species} \times \text{sex}$$



4. Interpret

```
Analysis of Variance Table

Response: bill_length_mm
             Df Sum Sq Mean Sq  F value Pr(>F)
species       2 7015.4  3507.7 654.1894 <2e-16 ***
sex           1 1135.7  1135.7 211.8066 <2e-16 ***
species:sex   2   24.5    12.2   2.2841 0.1035
Residuals   327 1753.3     5.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

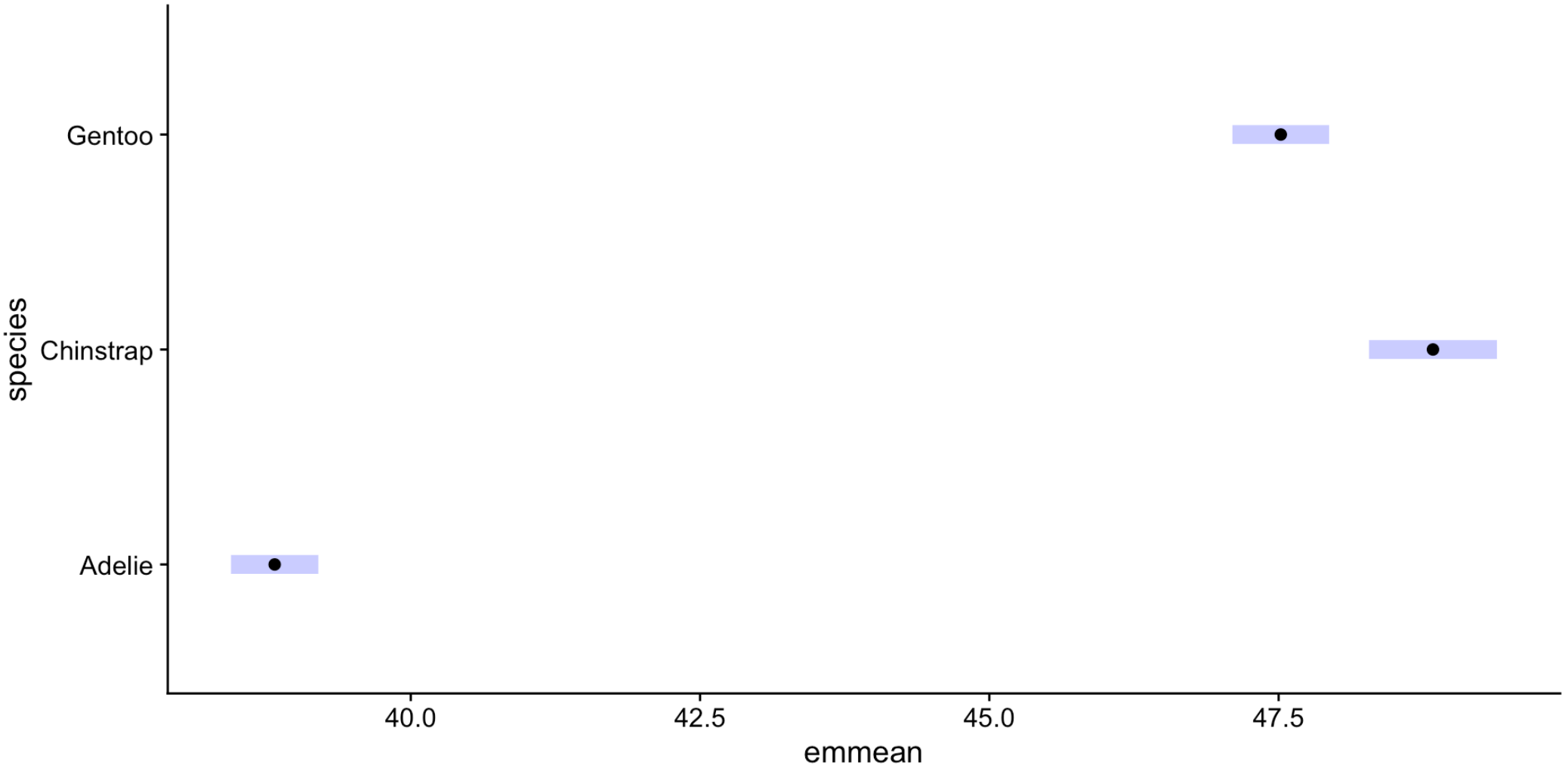
No significant interactions, **so we can interpret the main effects**.

- The results revealed significant main effects for both species and sex.
- Specifically, the effect of species on bill length was statistically significant, $F(2, 327) = 654.19$, $p < .001$, indicating that bill length varies significantly across different species.
- The effect of sex was also significant, $F(1, 327) = 211.81$, $p < .001$, suggesting that bill length differs between males and females.

4. Post-hoc

To further investigate where the differences lie, we can perform a post-hoc test. The most common post-hoc is the Tukey HSD test (honestly significant difference), which can be performed using estimated marginal means (often referred to as emmeans).

```
$emmeans
 species   emmean    SE  df lower.CL upper.CL
 Adelie      38.8 0.192 327     38.4     39.2
 Chinstrap   48.8 0.281 327     48.3     49.4
 Gentoo      47.5 0.212 327     47.1     47.9

Results are averaged over the levels of: sex
Confidence level used: 0.95

$contrasts
 contrast           estimate    SE  df t.ratio p.value
 Adelie - Chinstrap   -10.01 0.340 327 -29.444  <.0001
 Adelie - Gentoo       -8.69 0.286 327 -30.398  <.0001
 Chinstrap - Gentoo     1.32 0.352 327   3.735  0.0006

Results are averaged over the levels of: sex
P value adjustment: tukey method for comparing a family of 3 estimates
```



Post-hoc tests revealed that Chinstrap penguins had significantly longer mean bill lengths (48.8 mm ± 0.28 SE) compared to Adelie (38.8 mm ± 0.19 SE) and Gentoo penguins (47.5 mm ± 0.21 SE), and that all pairwise comparisons were statistically significant ($p < .001$).

Questions to consider

- When is it more appropriate to summarise a GLM with an ANOVA table versus regression coefficients?
- How do the standard GLM assumptions (LINE) apply when working with categorical predictors?
- How does the GLM formulation and interpretation change when including one vs. two categorical predictors, especially regarding main effects and interactions?
- What does a significant interaction term in a GLM tell us about the relationship between categorical predictors?
- Why are post-hoc tests necessary for interpreting GLMs with significant categorical predictors with more than two levels?

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.