Modelling multiple relationships: a linear model with multiple, continuous 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

Learning objectives

You should:

- [ ] Understand how to include multiple predictors in a model.
- [ ] Be able to interpret the coefficients of a multiple linear regression (MLR) model and report the results.
- [ ] Know how to assess assumptions of a MLR model.
- [ ] Understand the concept of multicollinearity and how to check for it in a MLR model.
- [ ] Be able to understand when to include interactions in a model and how to interpret them, including how to visualise interactions.

> “Life is really simple, but we insist on making it complicated.”

— Confucius

Expanding a model

We start with the *simple* model:

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

But, we know that there are other (possible) factors that may influence the body mass of a penguin.

$$Sex$$

$$Species$$

$$Bill\ length$$

It is rarely the case that a response variable is influenced by a *single* explanatory variable.

How do we include multiple predictors in a model?

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

$$y \sim x$$

We add more predictors to the model by including them in the equation:

$$y \sim x_1 + x_2 + x_3 + \ldots + x_n$$

where $x_1, x_2, x_3, \ldots, x_n$ are the predictors **that can be continuous or categorical** – or both.

Examples

- $\text{body mass} \sim \text{flipper length} + \text{bill depth}$ – all continuous
- $\text{body mass} \sim \text{flipper length} + \text{species}$ – at *least* one continuous, one categorical
- $\text{body mass} \sim \text{sex} + \text{species}$ – all categorical

Simple is best

- $\text{body mass} \sim \text{flipper length} + \text{bill depth}$ – all continuous
- $\text{body mass} \sim \text{flipper length} + \text{species}$ – at *least* one continuous, one categorical
- $\text{body mass} \sim \text{sex} + \text{species}$ – all categorical

The simplest model is one that is **ADDITIVE** – where in theory, each predictor contributes to the response variable *independently* of the other predictors.

Once the model is built, we can start to think about:

- data structure (e.g. continuous, categorical)– which defines the "traditional" name of the model – **Continuing…**
- interactions between predictors – **Week 4 (this week)**
- transformation to better fit the model – **Week 4 (this week)**
- the purpose of the predictors in the model (i.e. whether a predictor is considered a control, covariate, random variable, or fixed variable) – **Weeks 4 & 5**

In this lecture, we will focus on models with multiple predictors:

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

where all predictors are *continuous* variables – the multiple linear regression (MLR), and its interpretation.

Study design

- **Response variable**: body mass (g)
- **Predictors**: flipper length (mm) and bill depth (mm)
- **Control variables**: none

We define *no* control variables in this model, so both flipper length and bill depth are main effects of interest. More on control variables next week.

Why do we need multiple variables?

> The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

— John Tukey (1915-2000), *The Future of Data Analysis* (1962)

Controls and interactions

- **Control** – Add a variable which we believe **may influence the response variable**, but we are *not interested in its effect* on the response.
- **Interaction** – Because we added more variables, we want to assess how one variable influences the *relationship between the other variables*.

Depending on the research question(s), the above two reasons are common motivations for including multiple variables in a model.

> **Important**
>
> **Once interactions are included** the interpretation of the model then falls on either the:
>
> - **main effects** (interpretaition of the predictors that are of interest), or
> - **interactions** (interpretation of how predictors interact with each other)

Predictive power

- Include multiple variables **to increase the predictive power of the model**.
- **Focus** shifts – from *interpretation* to *prediction*.
- Anything goes – as long as the model "performs" better than before.
- **Not the focus of BEDA but important in other fields (e.g. machine learning).**

Important

Simple models are preferred in biological research, especially when data is collected from the field. Try not to overcomplicate the model and include additional variables only if they are necessary (controls or interactions) or part of the research question.

Additional considerations of MLR

Once we include multiple predictors in the model, the interpretation of the model changes:

1. **Coefficients** (β_i) are now **conditional** on the other predictors being fixed.
2. **Hypotheses** can be tested for each predictor via traditional p-values.
3. **Interactions** are now possible and should be considered when including multiple predictors. **If interactive effects are statistically significant, we place more emphasis on the interaction terms than the main effects.**
4. **Multi-collinearity** – the correlation between predictors – are considered if interactions are not included; otherwise ignore it if interactions are included and significant.
5. **Transformations** are more common in MLR due to the scale of the predictors. *This will be covered tomorrow.*

Example: penguins dataset

Starting with a simple linear regression model…

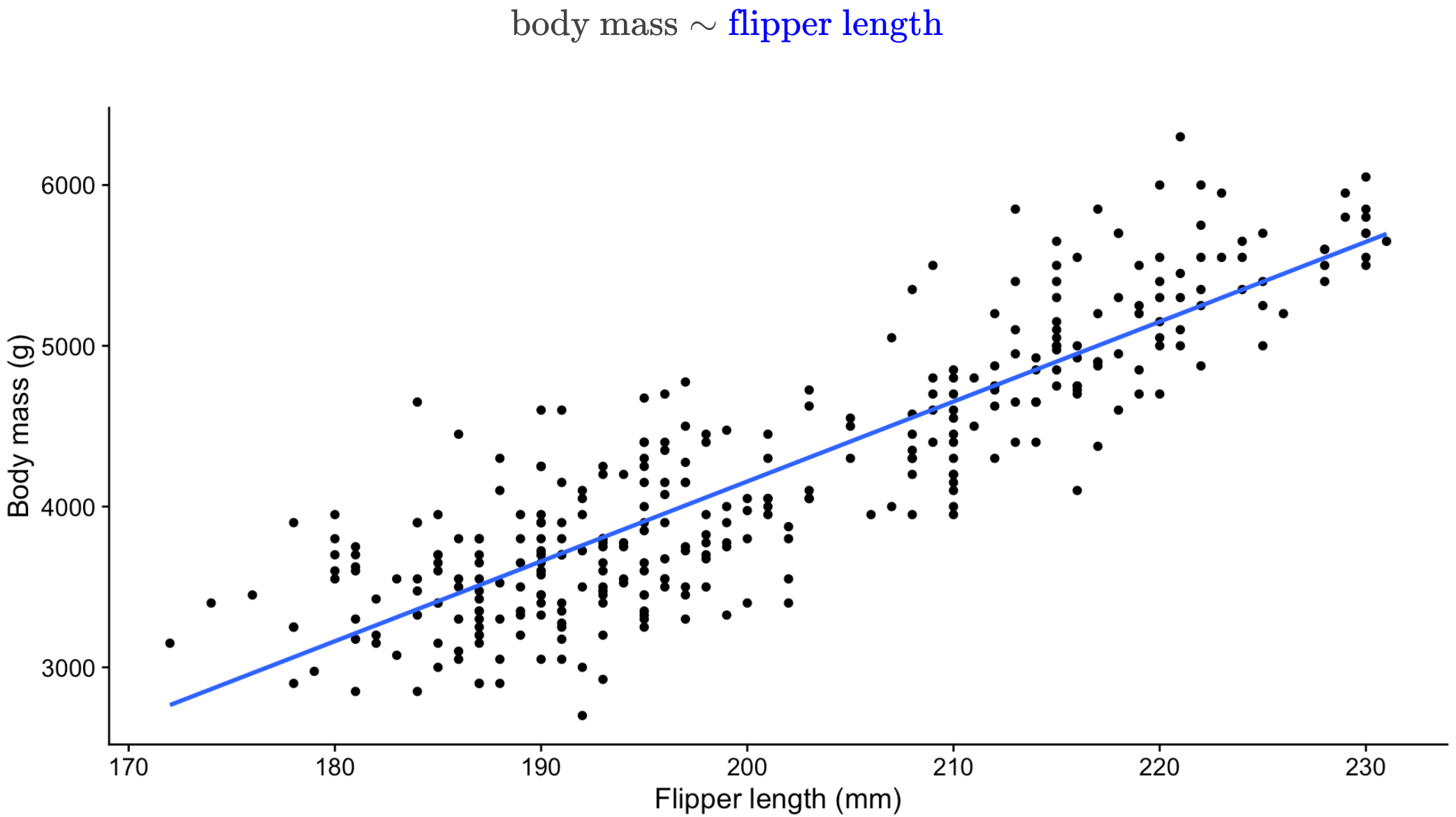
Simple linear regression

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

Four steps:

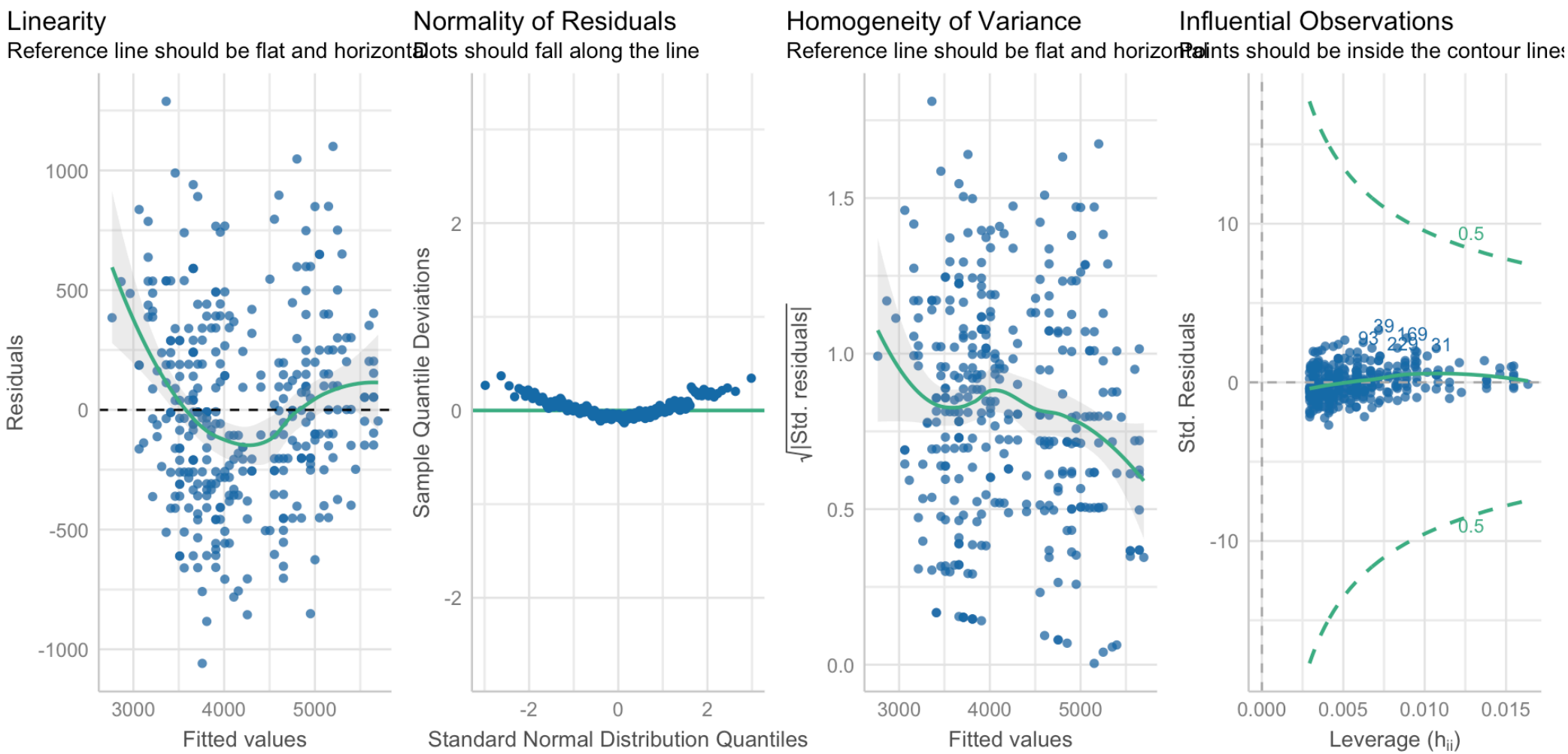
1. **Fit the model**, but don't interpret yet. Visualise the relationship (if possible).
2. **Check assumptions** from diagnostic plots (residuals).
3. Select a different model or transform data if assumptions are violated, go back to (2). Skip if assumptions are met.
4. **Interpret** the model.

1. Model/plot


body mass ~ flipper length
Body mass (g)
6000
5000
4000
3000
170
180
190
200
210
220
230
Flipper length (mm)

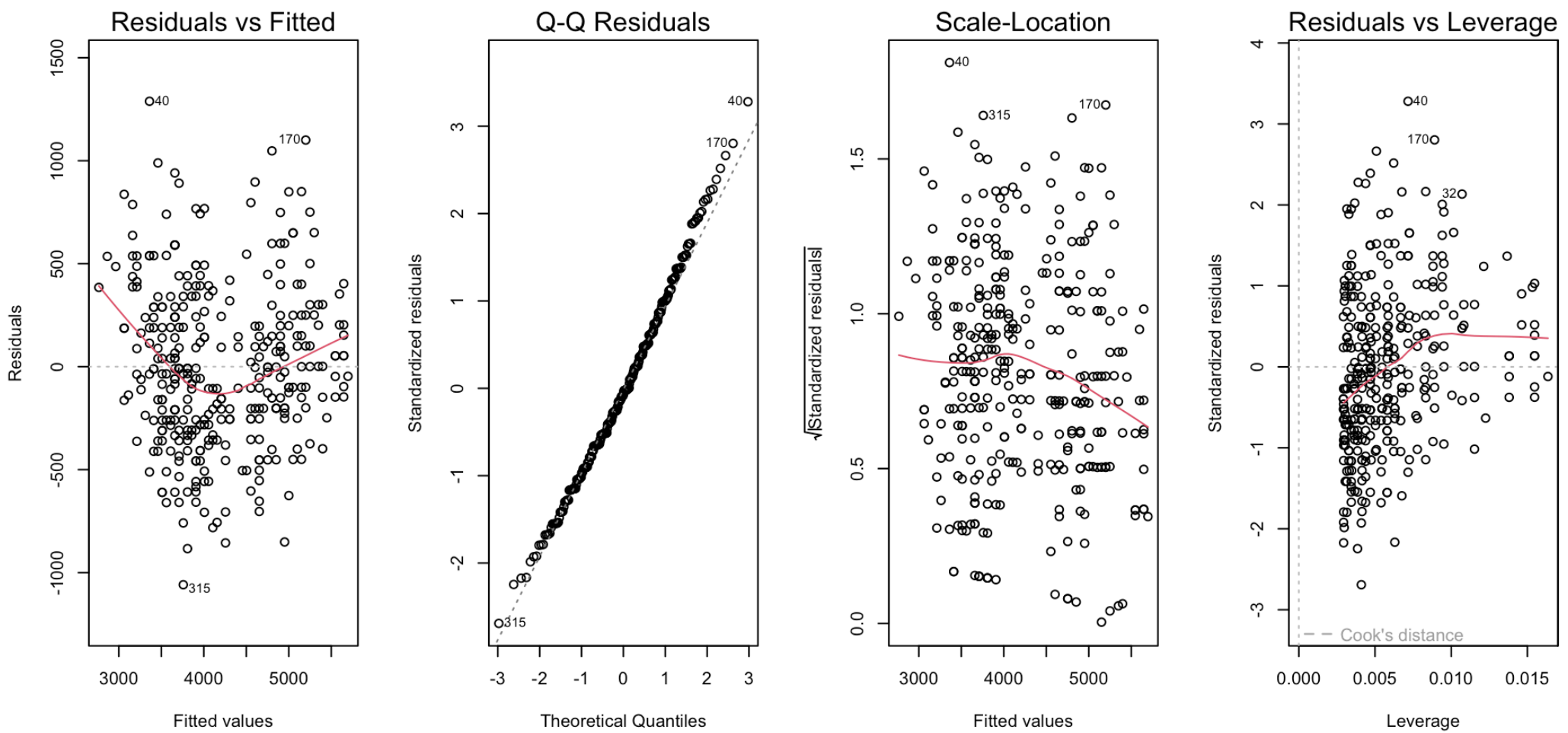
2. Diagnostic plots (assumptions) - `performance`

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



2. Diagnostic plots (assumptions)

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



4. Interpret

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

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max
-1058.80  -259.27   -26.88   247.33  1288.69

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5780.831    305.815  -18.90   <2e-16 ***
flipper_length_mm    49.686      1.518   32.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 394.3 on 340 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.759, Adjusted R-squared:  0.7583
F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

4. Interpret

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

Table 1: Summary table of regression analysis for predicting body mass from flipper length. Note: σ is the standard error of the residuals.

Characteristic	Beta	p-value
(Intercept)	-5,781	<0.001
flipper_length_mm	50	<0.001
R^2 = 0.759; σ = 394		

> The results of the multiple linear regression analysis indicate that the flipper length is a highly significant predictor of the body mass of penguins ($p < 0.001$, Table 1). The model demonstrates a strong relationship between the two variables, explaining 75% of the variance in the body mass ($R^2 = 0.76$). Specifically, the analysis reveals that for every 1 mm increase in flipper length, the body mass of a penguin increases by 50 g.

Adding more predictors

A *multiple* linear regression model

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

Four steps (it doesn't change!):

1. **Fit the model**, but don't interpret yet. Visualise the relationship (if possible).
2. **Check assumptions** from diagnostic plots (residuals), **including multicollinearity**.
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.

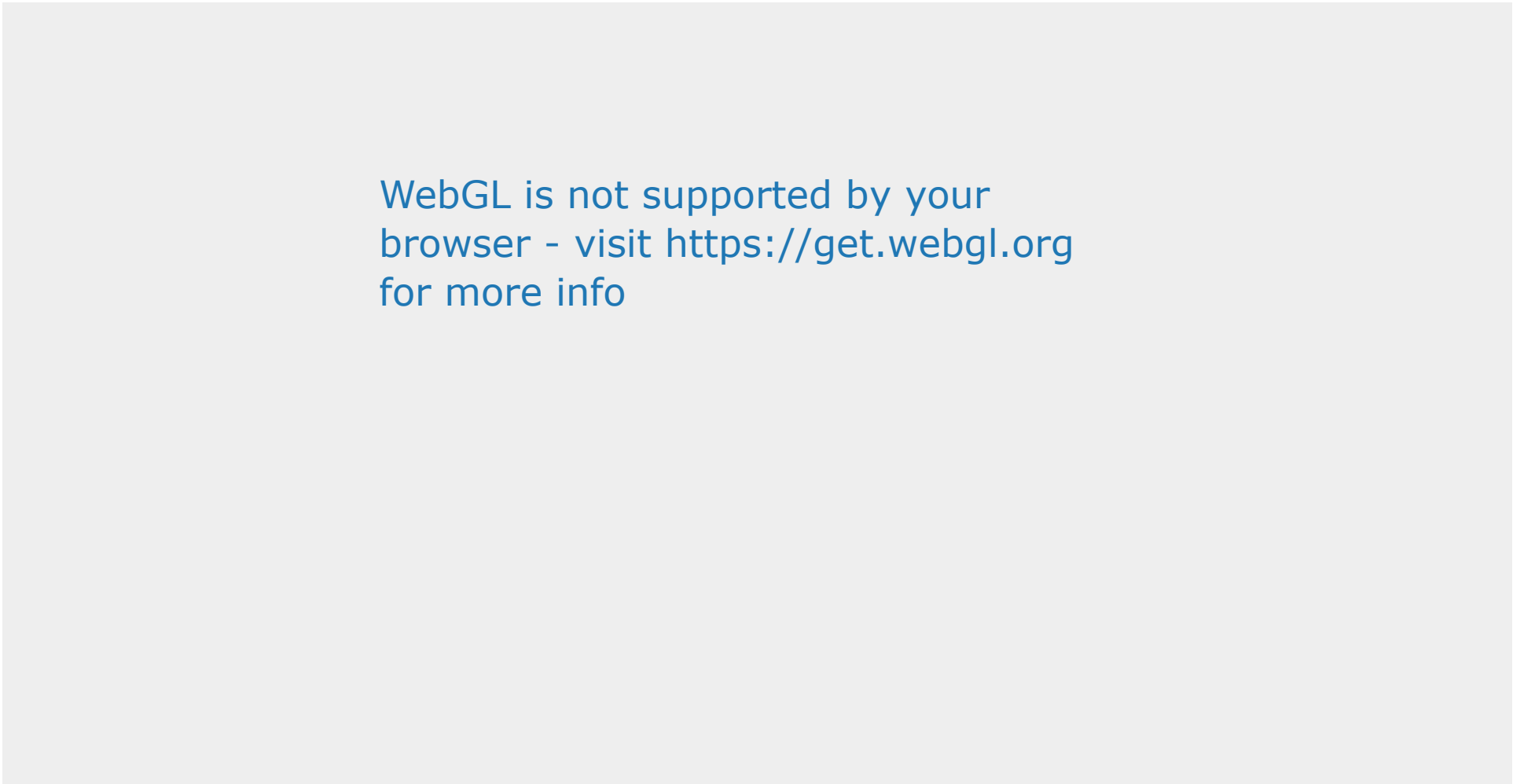
Multicollinearity

- **Multicollinearity** is the correlation between predictors in the model.
- **Problem**:
  - ➠ It can inflate the standard errors of the coefficients (i.e. make error margins wider), making the interpretation of the coefficients less reliable.
  - ➠ Interpretation of coefficients becomes difficult – since they are correlated, the effect of one predictor is confounded by the other predictor.
- **Solution**: Check the correlation between predictors and consider removing one of the predictors if the correlation is too high. Rule of thumb: $r > 0.7$.
  - ➠ Alternatively: check the variance inflation factor (VIF) – a measure of how much the variance of the coefficient is inflated due to multicollinearity. **A value of 5 or more is considered problematic.**

> **Note**
>
> Multicollinearity is not a problem if
>
> - **There is significant interaction.** We won't interpret the main effects anyway.
> - **The goal is prediction**. Although, it could affect computational efficiency.

1. Model/plot

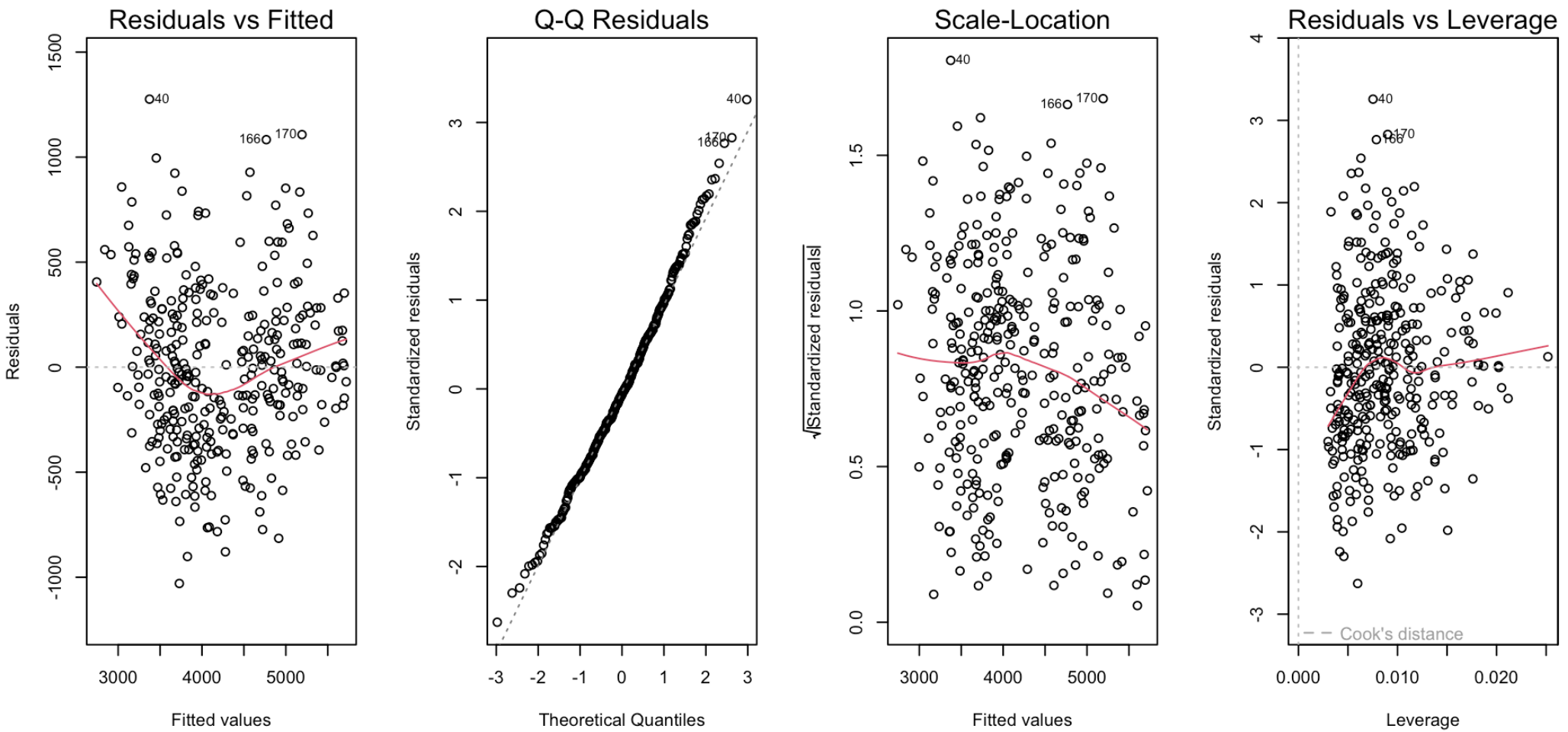
$$\text{body mass} \sim \text{flipper length} + \text{bill depth}$$



Visualising MLR is challenging – perhaps the maximum number of predictors we can visualise is 2 against the response variable. In most cases we rely on the diagnostic plots to interpret the model.

2. Diagnostic plots (assumptions)

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

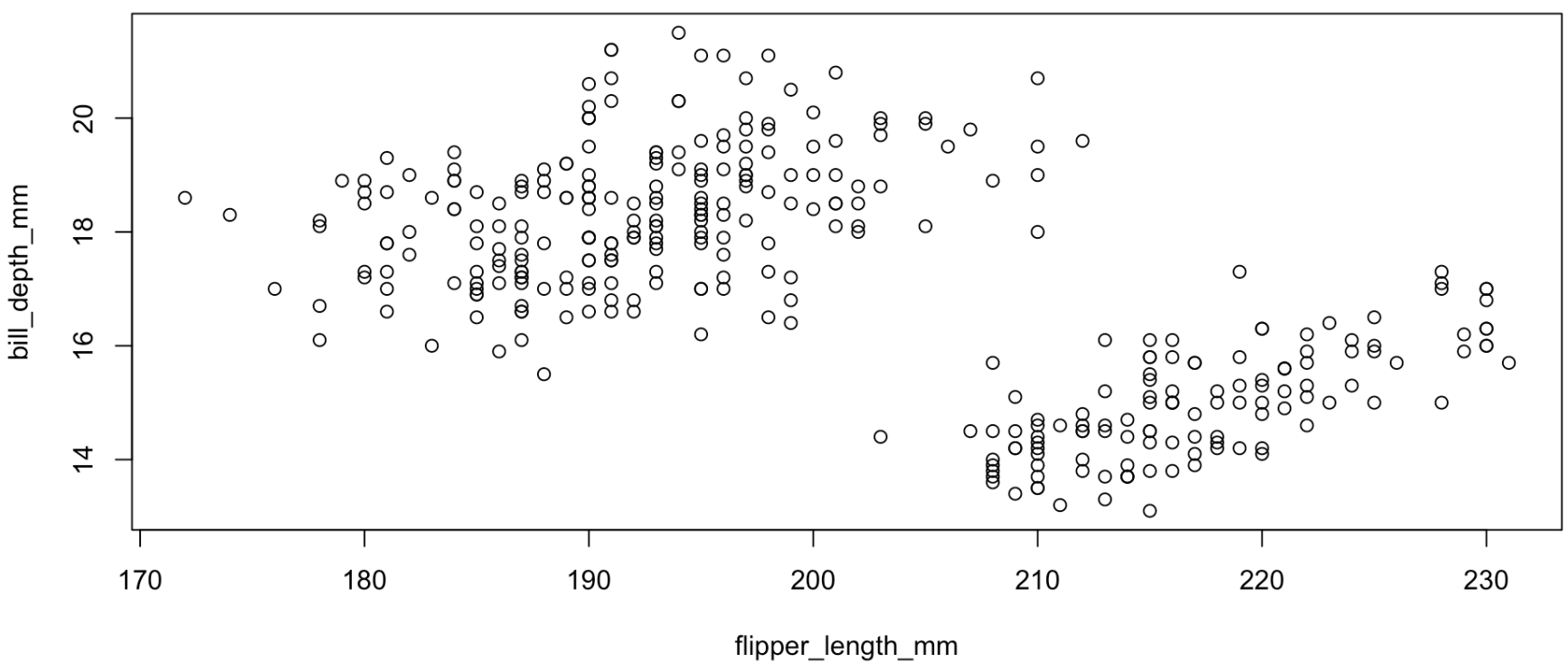


2. Multicollinearity

Correlation between predictors:

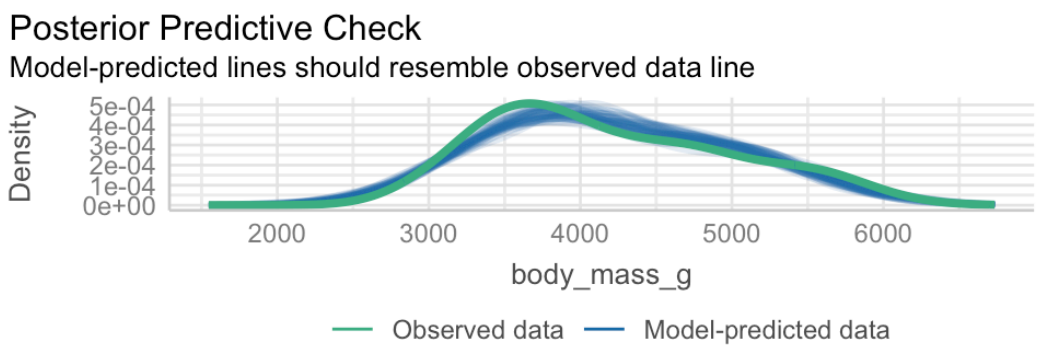
```
                  flipper_length_mm bill_depth_mm
flipper_length_mm                 1            NA
bill_depth_mm                    NA             1
```

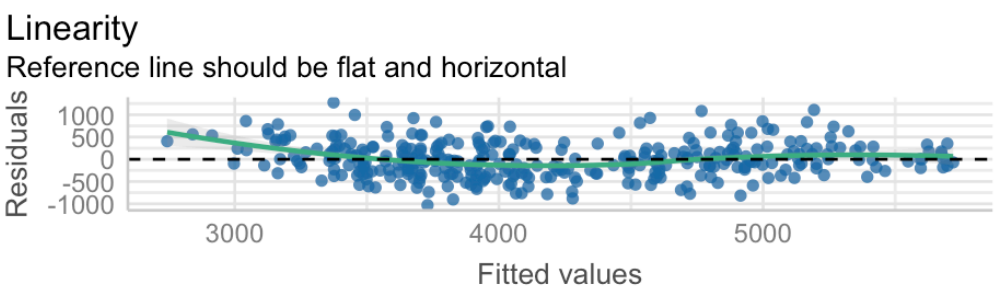
Plot the predictors:

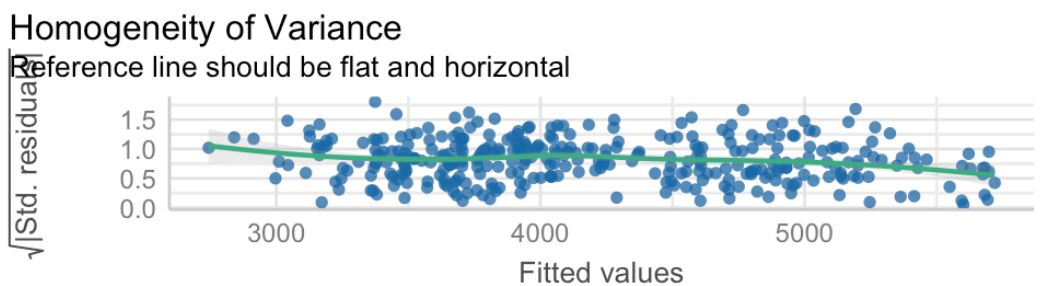


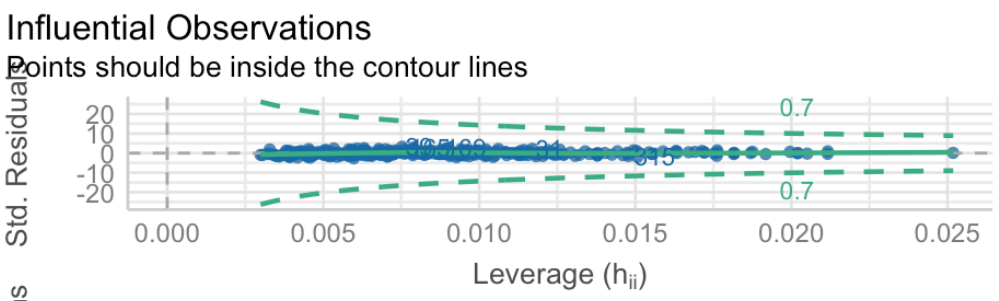
2. Multicollinearity

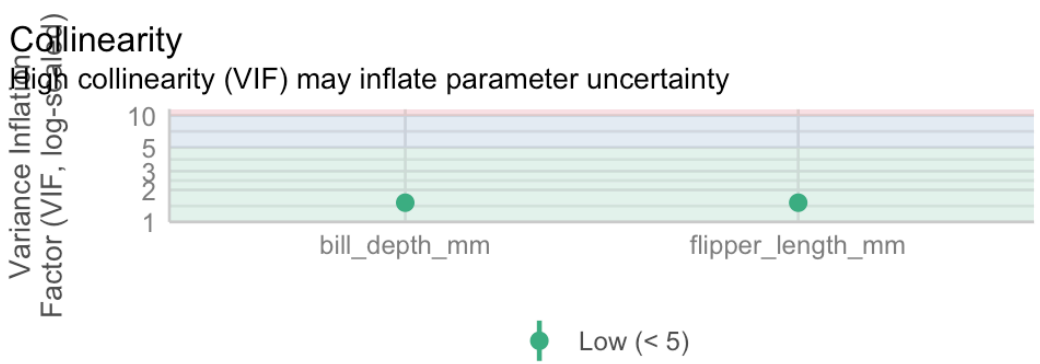
Or, use performance package and check the Collinearity Diagnostics:

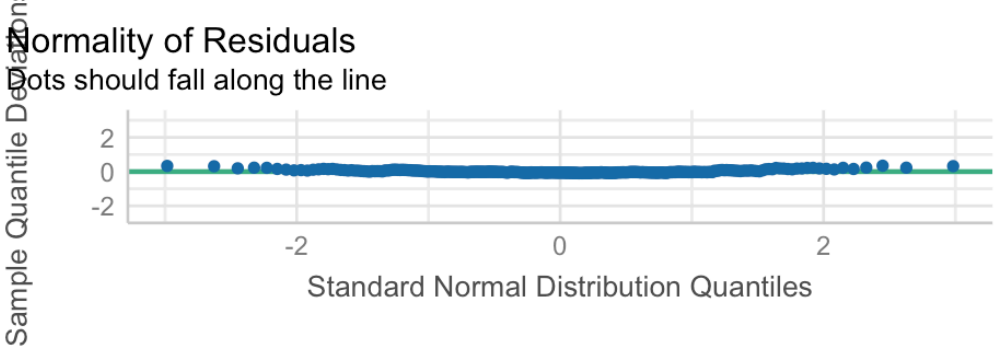












4. Interpret

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

Characteristic	Beta	p-value
flipper_length_mm	52	<0.001
bill_depth_mm	23	0.089
R^2 = 0.761; Adjusted R^2 = 0.760; σ = 393		

Interpretation (coefficients are now conditional)

- Flipper length is a significant predictor of body mass ($p < 0.001$), with a positive relationship where, for every 1 mm increase in flipper length, the body mass of a penguin increases by 52 g, *holding bill depth constant*.
- Bill depth is also a significant predictor of body mass ($p < 0.001$), with a positive relationship where, for every 1 mm increase in bill depth, the body mass of a penguin increases by 23 g, *holding flipper length constant*.
- Both predictors explain 76% of the variance in body mass (adjusted $R^2 = 0.76$).

Interactions

Will the relationship between A and B change, if C changes?

When to include interactions?

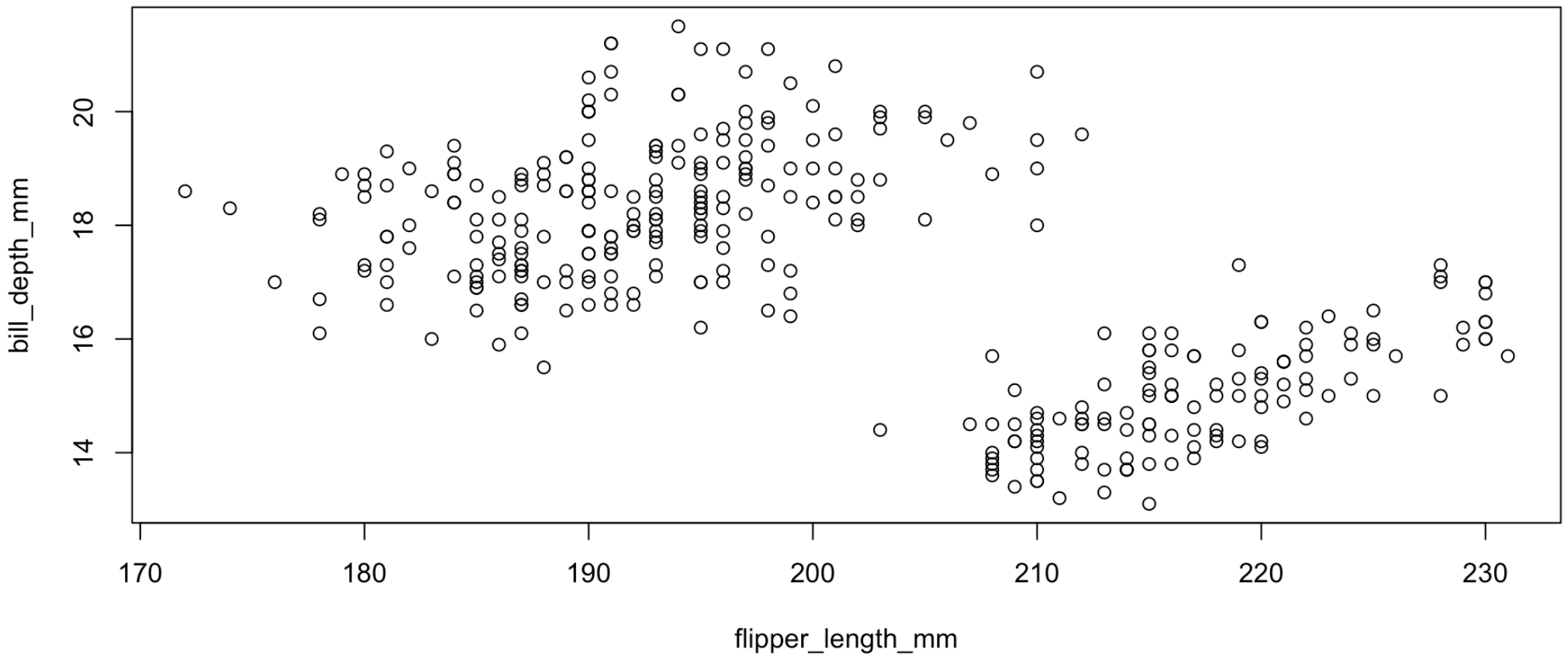
For any predictor that is included in the model and is not used as a control, we should consider interactions with other predictors.

Why interactions?

Interactions allow us to assess how one predictor influences the relationship between the response variable and another predictor. Examples:

- Does sunlight exposure affect plant growth? It **depends** on water availability – i.e. sunlight exposure and water availability may interact such that the effect of sunlight exposure on plant growth is different depending on the level of water availability.
- Does the effect of a drug on blood pressure depend on the age of the patient? It **depends** on the age of the patient – i.e. the effect of the drug on blood pressure may be different depending on the age of the patient.

Recall the plot of flipper length and bill depth



Do you think the two predictors are interacting?

Including interactions in the model

For a given MLR model

$$y \sim x_1 + x_2$$

we can declare an interaction term between x_1 and x_2 by changing the operator from $+$ to $\times$ or $*$:

$$y \sim x_1 \times x_2$$

So for the current MLR model from the pengins dataset:

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

becomes:

$$\text{body mass} \sim \text{flipper length} \times \text{bill depth}$$

which is also equivalent to:

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

Checking the model

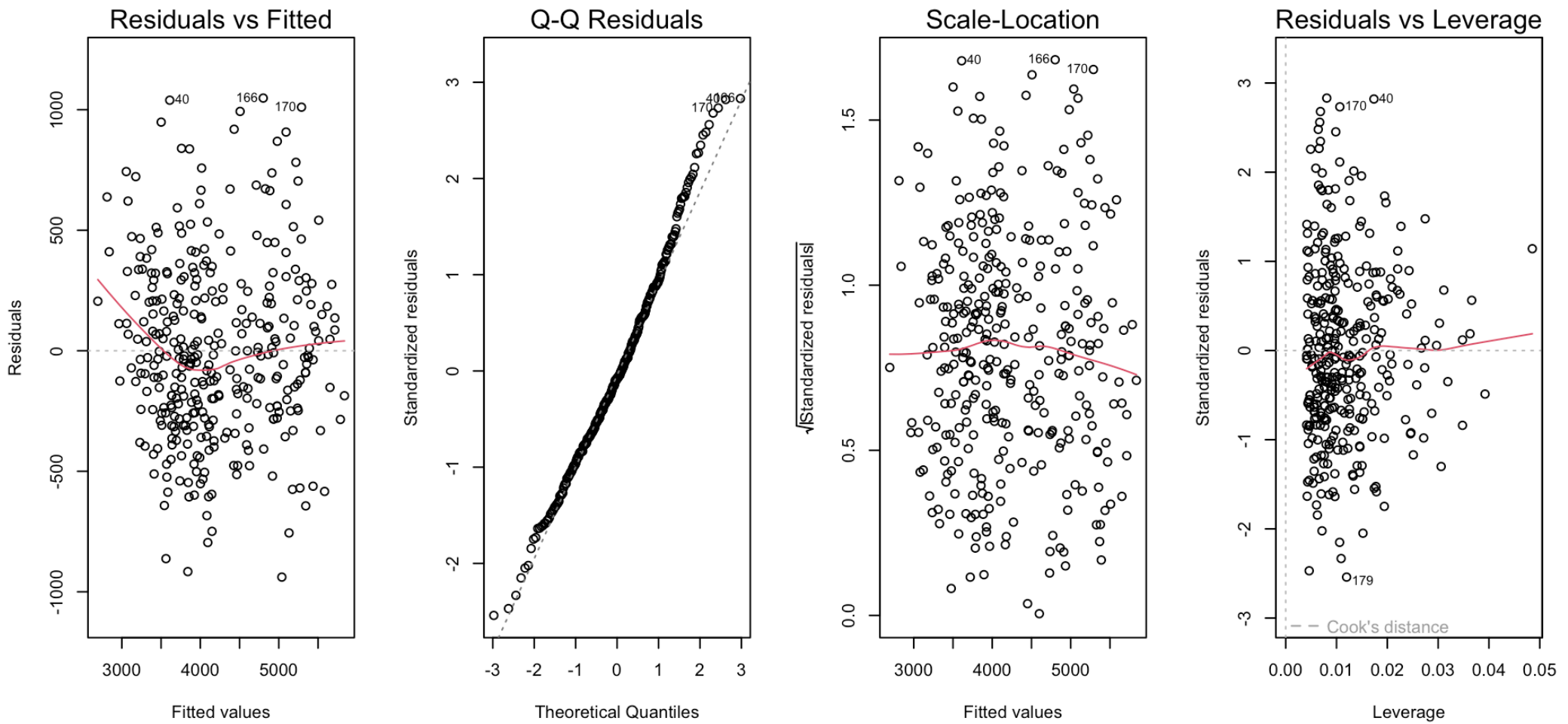
$$\text{body mass} \sim \text{flipper length} \times \text{bill depth}$$

Four steps (it doesn't change!):

1. **Fit the model**, but don't interpret yet. Visualise the relationship (if possible).
2. **Check assumptions** from diagnostic plots (residuals), **including multicollinearity**.
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.

2. Diagnostic plots (assumptions)

$$\text{body mass} \sim \text{flipper length} \times \text{bill depth}$$



4. Interpret

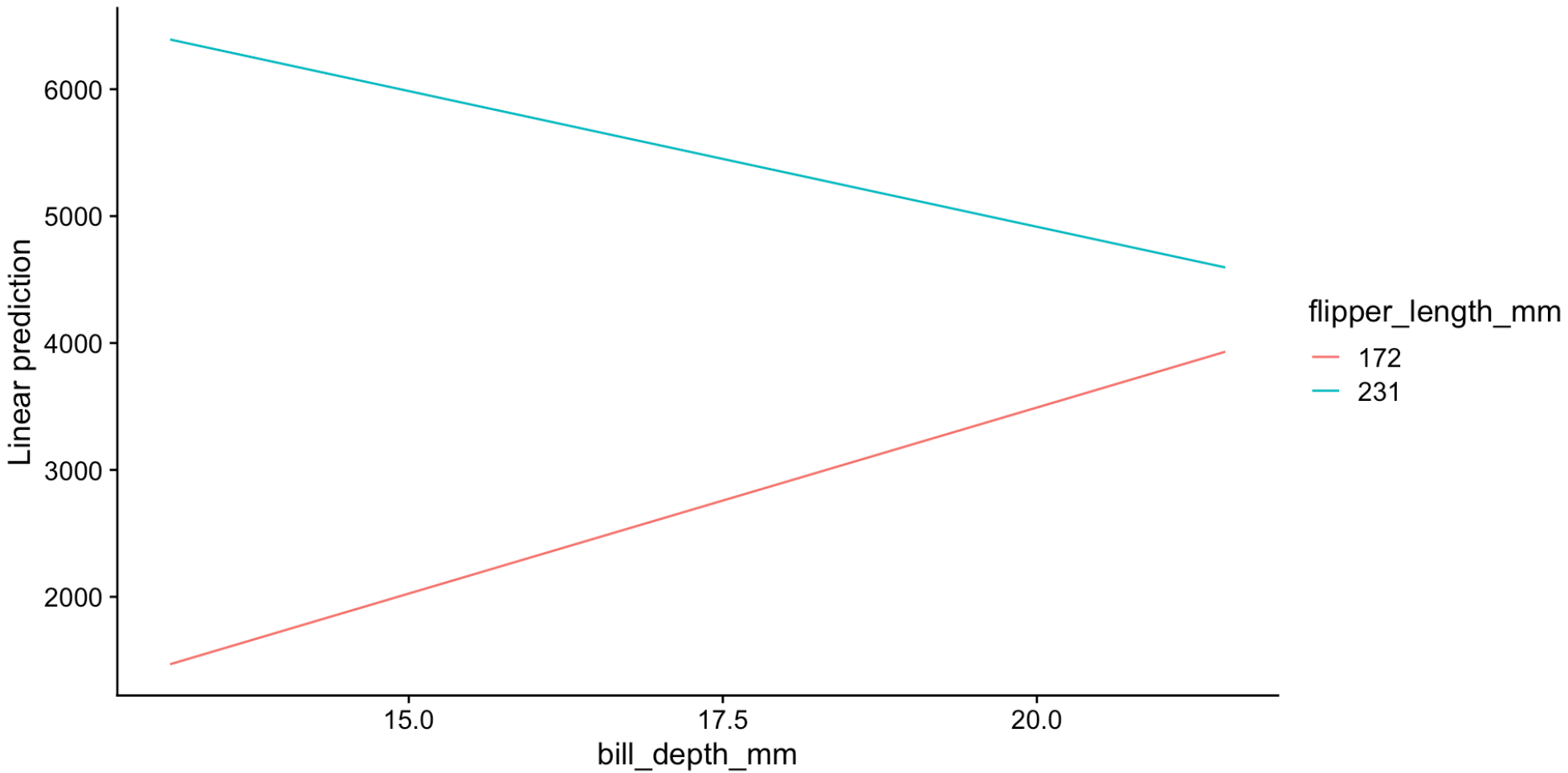
```
Call:
lm(formula = body_mass_g ~ flipper_length_mm * bill_depth_mm,
    data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-938.88 -253.96  -28.25  220.66 1048.33

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -36097.064   4636.271  -7.786 8.55e-14 ***
flipper_length_mm                196.074     22.603   8.675  < 2e-16 ***
bill_depth_mm                   1771.796    273.003   6.490 3.06e-10 ***
flipper_length_mm:bill_depth_mm    -8.596      1.340  -6.414 4.78e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 371.8 on 338 degrees of freedom
  (2 observations deleted due to missingness)
```

- If the interaction term is significant, it means that the interpretation of the main effects are most likely incorrect – we no longer interpret the main effects. **Instead, we figure out *how* those main effects are interacting.**

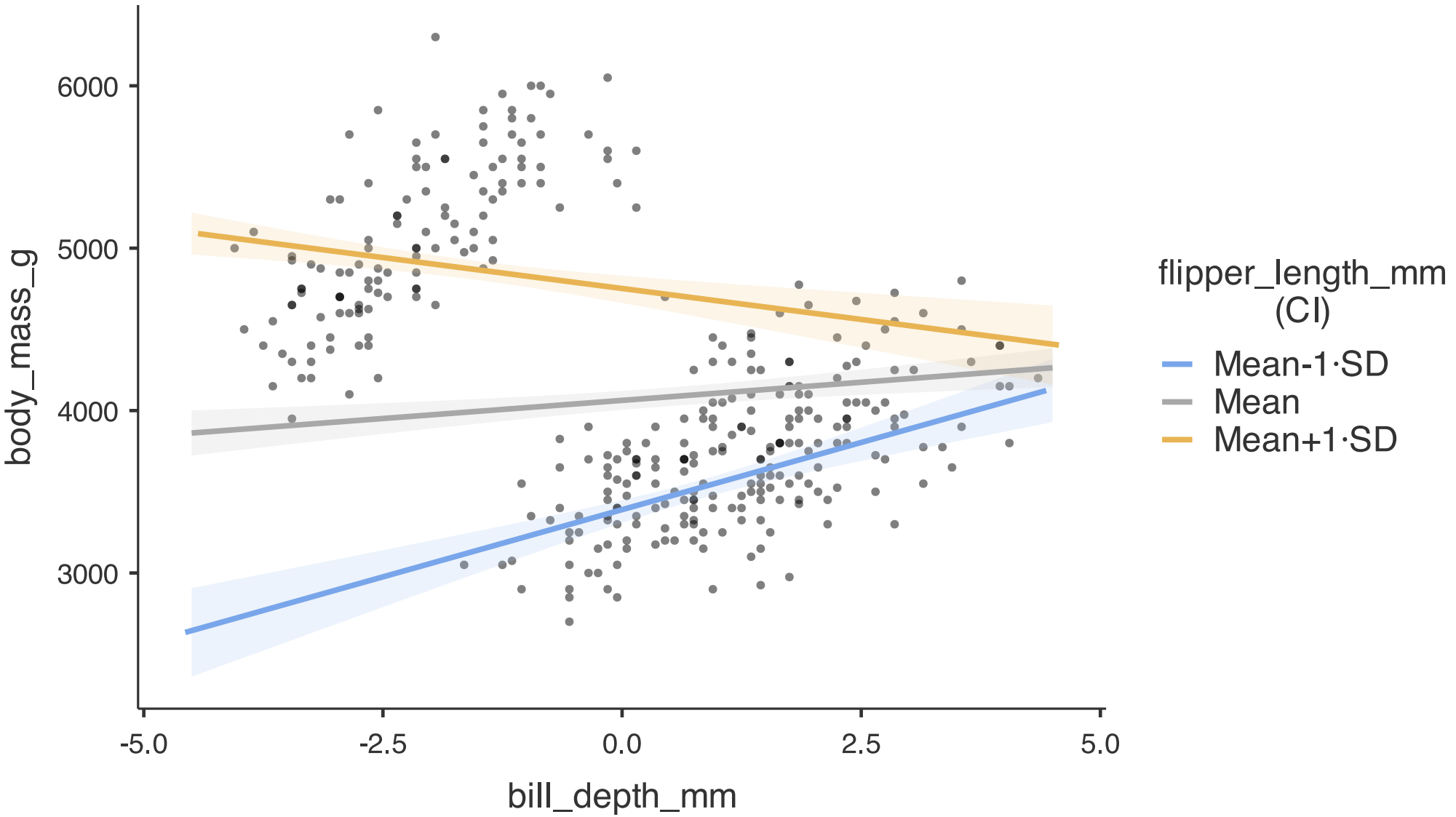
Interaction plots

To view the interaction, we need to work out at what “level” of a variable the relationship between the response variable and the other predictor is significant.

This can be automated in both R and Jamovi.



Interaction plot in Jamovi



Interpreting the plot

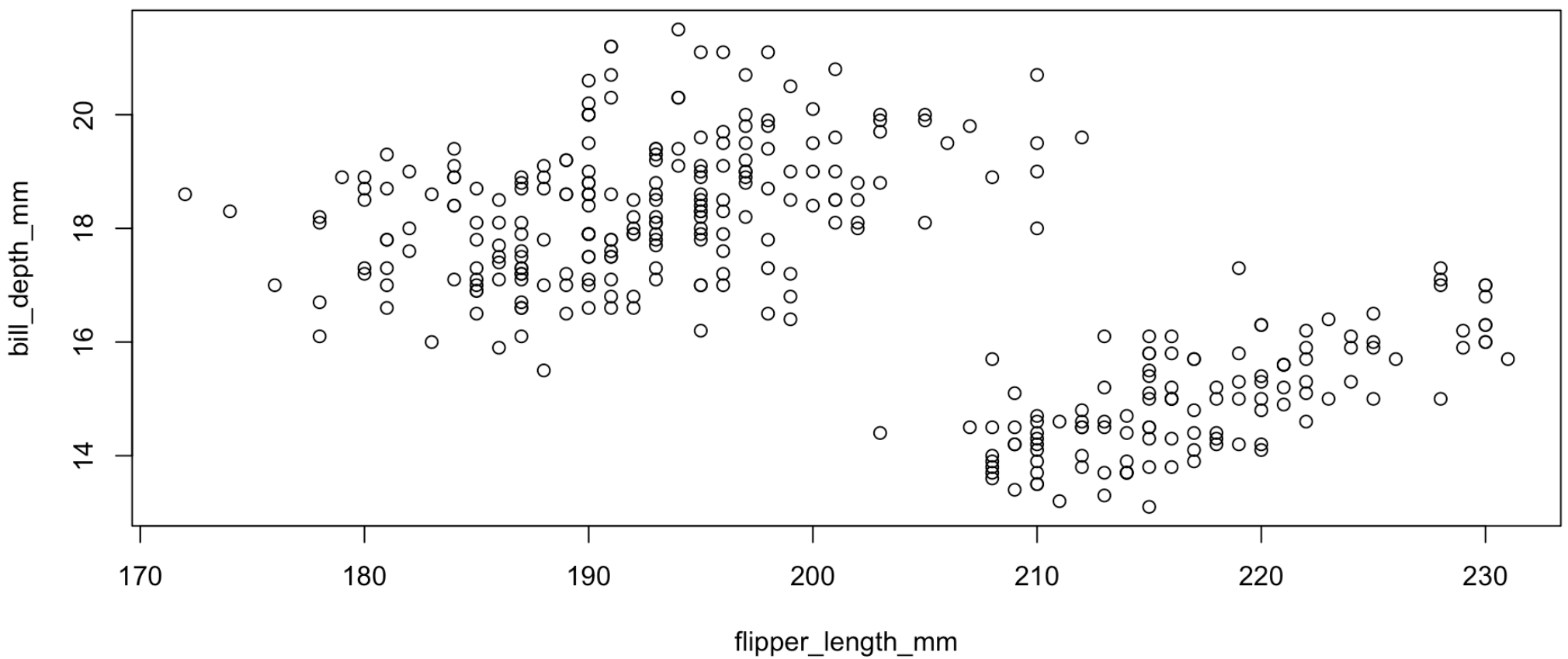
> Does bill depth affect the predicted outcome? It depends on flipper length!

Results indicate a significant interaction between flipper length and bill depth ($p < 0.001$). The interaction plot shows that the relationship between flipper length and body mass is dependent on bill depth. Specifically:

- For smaller flipper lengths (e.g. 172 mm), increasing bill depth is associated with an increase in body mass.
- For larger flipper lengths (e.g. 231 mm), increasing bill depth is associated with a decrease in body mass.

Note, R provides some information about what is “small” or “large” in the interaction plot, whereas Jamovi does not.

What is actually happening?



It is likely that something else is confounding the relationship between flipper length and body mass, as we have observed clustering in the plot of flipper length and bill depth.

Adding even more predictors: live demonstration

End of demonstration

Questions to consider

- What does it mean for a model to be additive?
- How do we interpret the coefficients of a multiple linear regression model?
- What are the assumptions of a multiple linear regression model?
- How do we check for multicollinearity in a multiple linear regression model?
- What is an interaction term in a multiple linear regression model, and how can we interpret it if it is significant?

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.