Forum 8
Data Skills for Scientists Regression: Multiple Numeric Predictor

2025-09-24

Introduction

Regression

In regression we have:

- one numeric response variable, and
- some number of predictor or explanatory variables that can be numeric or categorical.

Regression

This third and final module, 'Regression' contains three topics, in which we cover regression depending on which kinds of predictors are used.

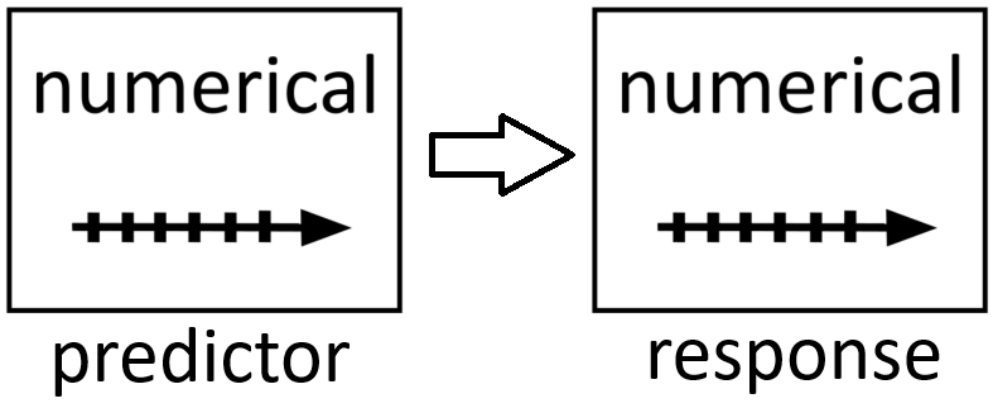
In particular:

- One Numeric Predictor;
- Multiple Numeric Predictors; and
- Categorical Predictors.

This week, we will cover the second of these topics — Multiple Predictors.

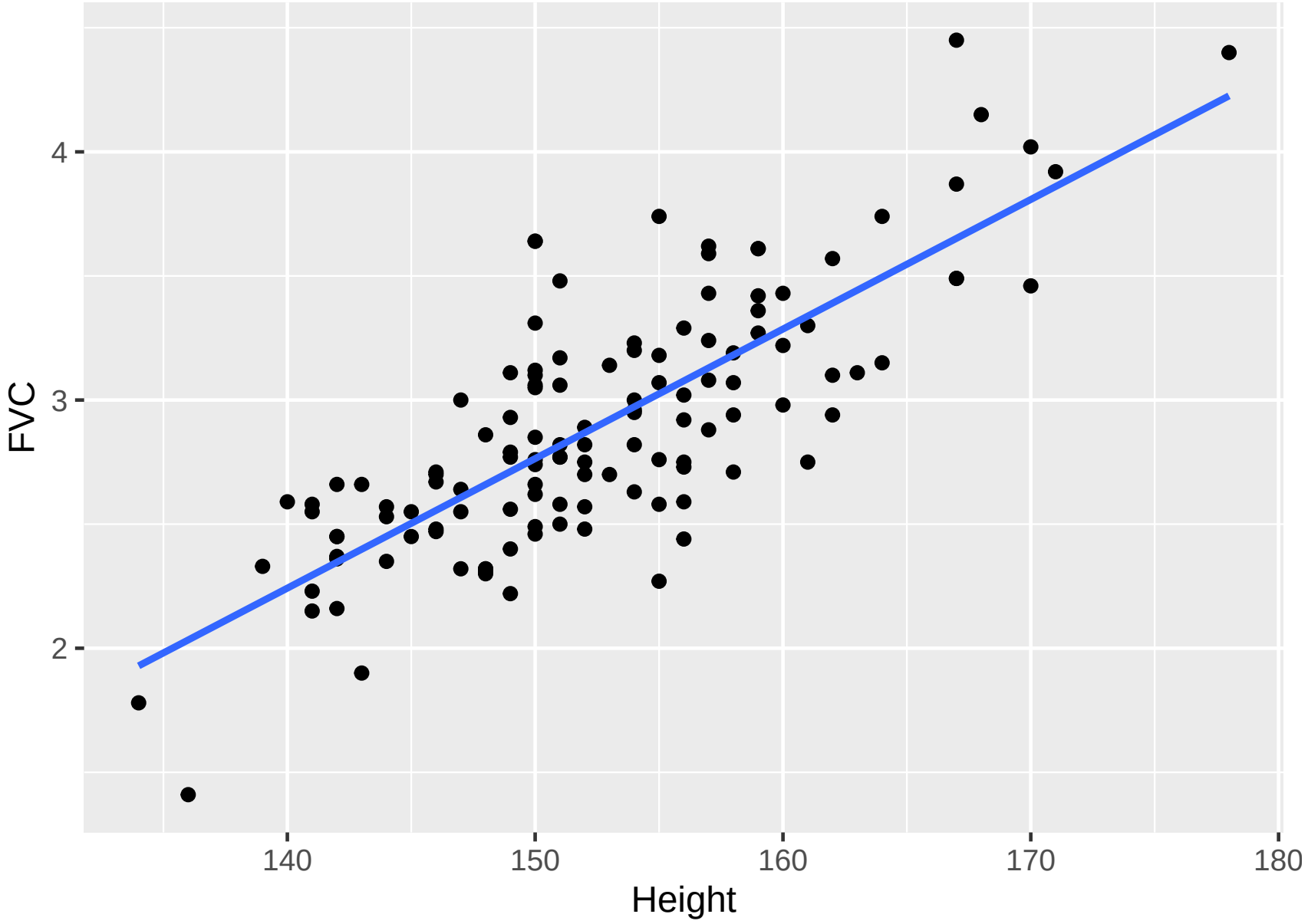
One Numeric Predictor

In the previous topic/ forum, we introduced regression with one numeric predictor. This is often called 'simple' linear regression.



Recall the `FVC` data?

Example - FVC versus Height



Topic Summary

We will extend the 'simple' linear regression concept to including multiple predictors in our linear regression — something often called 'multiple' linear regression. In particular:

- Two numeric predictors, but then we will also introduce using
- One numeric and one categorical predictor (with and without interaction terms).

Sub-topics

Much like in the previous forum, in todays forum we will cover:

- Fitting a regression line to data,
- Checking model assumptions,
- Inference for regression parameters, and
- Making predictions.

For each of the three cases we'll introduce.

Other Cases

It is possible to fit a multiple linear regression model with:

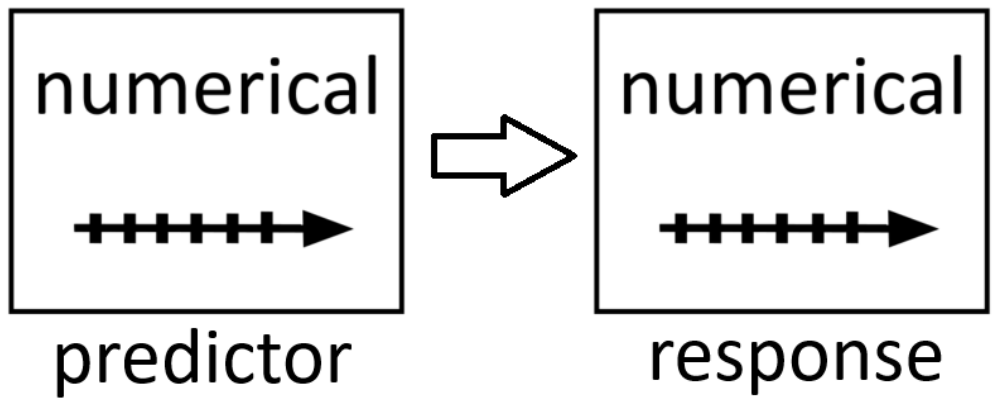
- several categorical predictors, and
- multiple numeric predictors with interaction terms,

But we won't cover those cases in this course. If you're interested, you can explore these cases for yourself based on what we cover in this forum.

Two Numeric Predictors

Linear Regression on One Numeric Predictor

Recall we introduced 'simple' linear regression on one numeric predictor



with the equation for the regression line

$$y = b_0 + b_1 x$$

and used the FVC data to illustrate it by fitting a model to predict `FVC` based on `Height`?

Example - FVC versus Height

```
Call:
lm(formula = FVC ~ Height, data = fvc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75507 -0.23898 -0.00411  0.21238  0.87589

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.064961   0.552593  -9.166 1.24e-15 ***
Height       0.052194   0.003618  14.426  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3137 on 125 degrees of freedom
Multiple R-squared:  0.6248,    Adjusted R-squared:  0.6218
F-statistic: 208.1 on 1 and 125 DF,  p-value: < 2.2e-16
```

Example - FVC versus Height

Where our estimate for the intercept and slope were approximately $b_0 = -5$ and $b_1 = 0.05$ respectively. So our estimated regression line was approximately

$$y = b_0 + b_1 x = -5 + 0.05x$$

Note that the R-squared or coefficient of determination for this model is 0.62 indicating that this model can explain 62% of the variability in `FVC` based on using variation in `Height` through the fitted regression line.

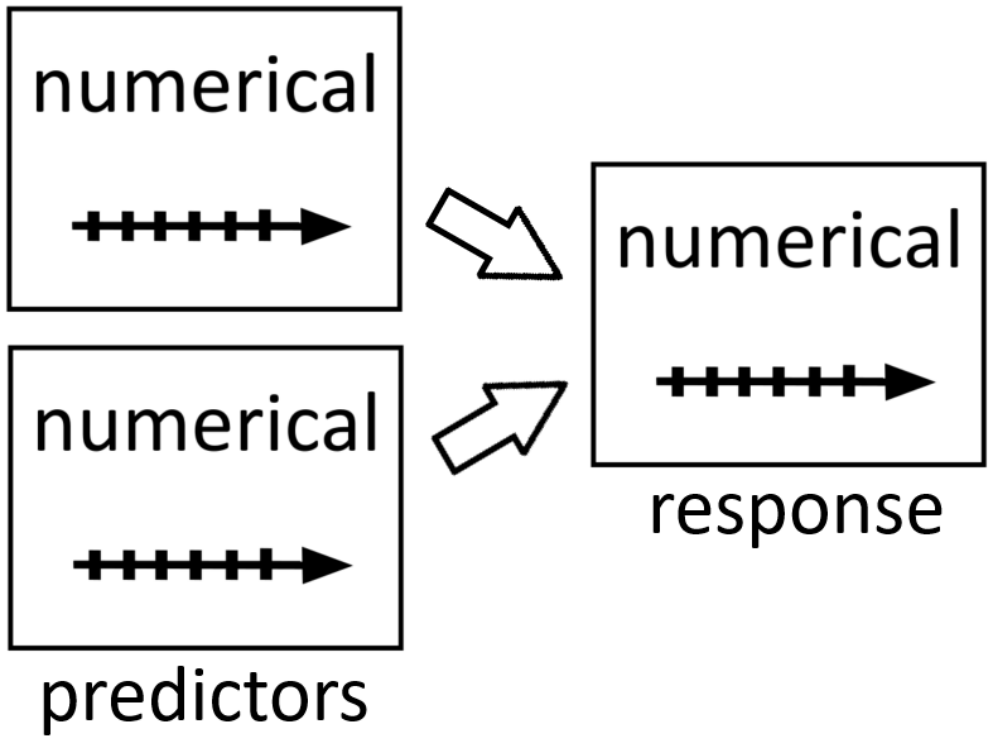
Example - FVC - Fit Summary

Recall the FVC data has another variable recorded — weight measured in kilograms — as well as FVC (in litres) and height (in centimetres) which we used in the model we explored.

> *Could we add `Weight` to our model and use it in combination with `Height` and get even better predictions of `FVC`?*

Two Numeric Predictors

We can add another numeric predictor to our regression model and use both predictors!



Two Numeric Predictors

This corresponds to adding another term to the equation to include the second numeric predictor, so the equation becomes

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

where x_1 and x_2 represent our two predictor variables.There is now an extra 'slope' coefficient b_2.

The interpretation of all three coefficients b_0, b_1 and b_2 changes in comparison to the 'simple' linear regression model with only one numeric predictor.

This is because the regression equation now no longer represents a 'line', but we don't need to worry about this too much.

Interpretation of b_0

b_0 is still an intercept, but in comparison to the model with one numeric predictor in which the interpretation of b_0 was

If the <explanatory variable> is zero, then we expect the mean value of the <response variable> to be $<b_0>$ <units>.

Now, the interpretation of b_0 becomes

If **both** <explanatory variables> are zero, then we expect the mean value of the <response variable> to be $<b_0>$ <units>.

Interpretation of b_1 and b_2

Both b_1 and b_2 are slopes, in a way. However in comparison to the model with one numeric predictor in which the interpretation of b_1 was

If the $<x_1>$ increases by one <unit>, then we expect the mean value of the <response variable> to <increase/decrease> by $<|b_1|>$ <units> on average.

Now, the interpretation of b_1 becomes

If the $<x_1>$ increases by one <unit> **and** x_2 **remains unchanged** we expect the mean value of the <response variable> to <increase/decrease> by $<|b_1|>$ <units>.

and similarly, the interpretation of b_2 will be

If the $<x_2>$ increases by one <unit> **and** x_1 **remains unchanged** we expect the mean value of the <response variable> to <increase/decrease> by $<|b_2|>$ <units>.

Example - FVC versus Height and Weight

```
Call:
lm(formula = FVC ~ Height + Weight, data = fvc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65960 -0.21612 -0.00273  0.17748  0.88240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.799689   0.664114  -5.721 7.48e-08 ***
Height       0.039651   0.005252   7.549 8.23e-12 ***
Weight       0.014871   0.004652   3.197  0.00176 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3027 on 124 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6477
F-statistic: 116.8 on 2 and 124 DF,  p-value: < 2.2e-16
```

Example - FVC versus Height and Weight

So the equation for our fitted regression model is approximately:

$$\text{FVC} = -3.8 + 0.04 \times \text{Height} + 0.015 \times \text{Weight}.$$

The coefficient of determination (R-squared) for this model is approximately 0.65 indicating that this model can explain 65% of the variability in FVC based on using variation in both Height and Weight through the fitted regression model. This is more than 0.62 when using only Height earlier.

> **Note**
>
> The R-square value will only ever increase (or stay the same) when adding additional explanatory variables to a linear regression model.

The adjusted R-square value is a little different, but no need to be concerned with that detail for now.

The Kitchen Sink

When choosing which variables to include and exclude in a regression model there is a tension between model fit (measured by R-squared value), and parsimony or simplicity.

Adding more variables improves the model fit (increases its R-squared value), but also adds complexity to the model. Simpler models are considered to be better as they are more parsimonious, essentially by Occam's Razor[1].

"Kitchen Sink Regression"[2] describes ignoring the value of parsimony and adding "everything but the kitchen sink" to a regression model in order to artificially inflate the R-square value. Generally speaking, this is bad practice.

---

[1] https://en.wikipedia.org/wiki/Occam%27s_razor.

[2] https://en.wikipedia.org/wiki/Kitchen_sink_regression.

Example - Interpretation of b_0

The 'intercept' coefficient estimate $b_0 = -3.8$ can be interpreted as predicting that a individual with zero height and weight would have a mean FVC of $b_0 = -3.8$ litres.

This is nonsense, since it is extrapolating to unreal individuals.

Example - Interpretation of b_1 and b_2

The 'slope' intercept $b_1 = 0.04$ indicates that if height increases by one centimetre but weight remains the same we would expect the mean FVC to increase by 0.04 litres.

In contrast, the second 'slope' intercept $b_2 = 0.015$ indicates that if **weight** increases by one kilogram but **height** remains the same we would expect the mean FVC to increase by 0.015 litres.

Fitted Values and Residuals - One Numeric Predictor

Consider the equation for the regression line

$$y = b_0 + b_1 x$$

and a data point (x_i, y_i). The **fitted value** for this data point is the y coordinate of the point on the regression line with the same x value as the data point x_i

$$\hat{y}_i = b_0 + b_1 \ x_i$$

and the **residual** or error for this data point is

$$r_i = y_i - \hat{y}_i,$$

Fitted Values and Residuals - Two Numeric Predictors

So similarly consider the equation for the regression model

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

and a data point (x_{i1}, x_{i2}, y_i). The **fitted value** for this data point is the y coordinate of the point on the regression model with the same x_1 and x_2 values as the data point x_{i1} and x_{i2}.

$$\hat{y}_i = b_0 + b_1 \ x_{i1} + b_2 \ x_{i2}$$

and the **residual** or error for this data point is

$$r_i = y_i - \hat{y}_i,$$

Example - Fitted Values and Residuals - FVC

Here are the first four rows of the FVC data:

FVC	Height	Weight
2.75	156	51
2.66	142	37
2.32	148	35
4.40	178	58

Let's consider the fourth data point, i.e. $i = 4$. So $x_{i1} = 178$, $x_{i2} = 58$, and $y_i = 4.4$.

Calculated the fitted value $\hat{y}_4$ and residual r_4 for this data point.

Example - Fitted Values and Residuals - Solution

Recall our fitted regression model was approximately

$$\text{FVC} = -3.8 + 0.04 \times \text{Height} + 0.015 \times \text{Weight}.$$

So the fitted value for this data point would be

$$\hat{y}_4 = -3.8 + 0.04 \times 178 + 0.015 \times 58 = 4.121$$

and the **residual** or error for this data point is

$$r_4 = 4.4 - 4.121 = 0.279$$

Assumptions of a Linear Regression Model

Recall that the assumotions for a linear regression model on one numeric predictor were:

- Linearity.
- Constant spread of residuals.
- Normality of the residuals.
- Independence of the residuals.

Although the meaning of 'linear' now that we have two numeric predictors is a little different, the assumptions for a linear regression model with two numeric predictors, and the way we assess them is broadly the same!

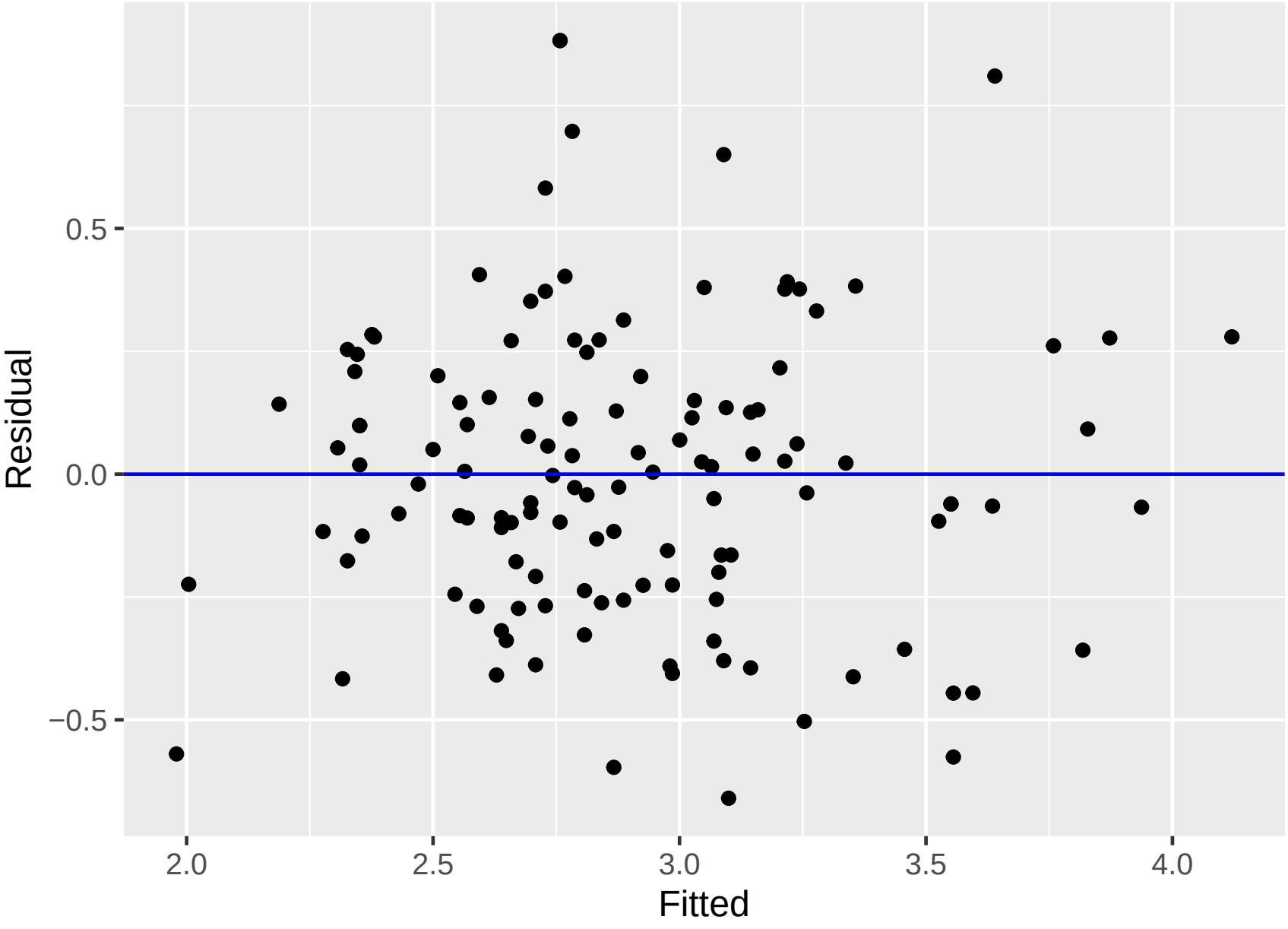
Plots to Check Assumptions

Just like with one numeric predictor, we assess three of the four assumptions with the same plots constructed from the fitted values and residuals

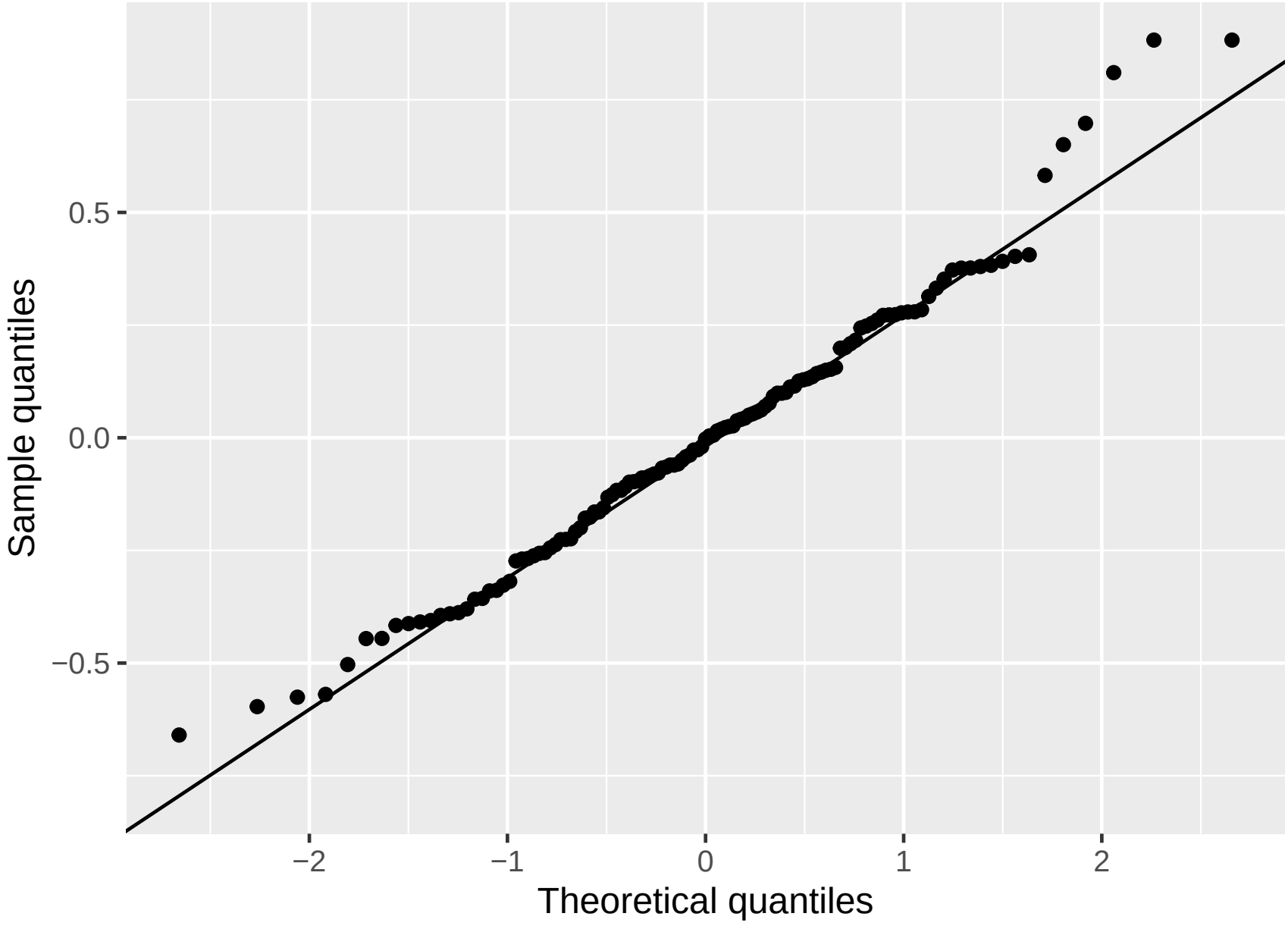
- Fitted vs Residuals, and
- Normal Quantile-Quantile of Residuals

that we used in the previous topic with one numeric predictor.

Example - Fitted versus Residuals Plot



Example - Normal QQ Plot of Residuals



Example - Check Assumptions

Are the

- Linearity.
- Constant spread of residuals.
- Normality of the residuals.

assumptions for this model reasonable?

Inference and Parameters - One Numeric Predictor

Recall we used b_0 and b_1 in the fitted regression line

$$y = b_0 + b_1 x$$

to estimate β_0 and β_1 in the population regression line

$$y = \beta_0 + \beta_1 x$$

and made inferences for the population parameters β_0 and β_1 with

- confidence intervals and
- hypothesis tests (t-tests)

We can do the same for a model with two numeric predictors.

Inference and Parameters - Two Numeric Predictors

We use b_0, b_1 and b_2 in the fitted regression model

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

to estimate the population parameters β_0, β_1 and β_2 in the population regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Warning

b_0 and b_1 in the fitted one numeric predictor model **are not the same** as b_0 and b_1 in the fitted two numeric predictor model (see the FVC example earlier). Similarly, β_0 and β_1 **are also not the same** in the two population models.

Example - FVC versus Height and Weight

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.7997	0.6641	-5.7214	0.0000
Height	0.0397	0.0053	7.5492	0.0000
Weight	0.0149	0.0047	3.1967	0.0018

To illustrate,

- Calculate a 95% confidence interval and
- Perform a t-test at a 1% significance level

for β_2.

Example - Confidence Interval for β_2

The critical value $t^* = 1.98$, $b_2 = 0.015$, and $SE_{b_2} = 0.0047$ (approximately) and we calculate the 95% confidence interval for β_2 as

$$b_2 \pm t^* \times SE_{b_2}$$

which is

$$0.015 \pm 0.009 \quad \text{or} \quad (0.006, 0.024)$$

and means we are 95% confident that for every kilogram that that weight increases while height is kept constant, we expect the mean forced vital capacity (FVC) to increase between 14 and 15 millilitres (mL).

Hypothesis Test for $\beta_2 = 0$

The hypotheses of our test for β_2 in the two numeric predictor model will be very similar:

$$H_0 : \beta_2 = 0,$$
$$H_a : \beta_2 \neq 0.$$

and it is used to answer the same question as in the one numeric predictor model:

> *"Does the explantory variable help predict the response variable?"*

Example - Hypothesis Test for β_2

The test statistic is

$$T = \frac{b_2}{SE_{b_2}} = \frac{0.015}{0.0092} = 3.2$$

The P-value of 0.00176 as shown in the summary results is obtained from the t-distribution with $n - 3 = 124$ degrees of freedom.

So we reject the null hypothesis and conclude that there is significant evidence that weight helps predict FVC in this model, at the 1% significance level since the P-value is less than 0.01.

Prediction

We can also use the two numeric predictor model to make predictions in very much the same way we used the one numeric predictor model, with

- Confidence intervals for the mean response value and
- Prediction intervals for a particular response value

given known values of **both explanatory variables**.

Example - Prediction Intervals

Recall that using the one numeric predictor model we predicted with 95% certainty that an individual with height 150 cm would have an FVC between 2.14L and 3.39L.

Lets use the two numeric predictor model now to see how the predictions we get for individuals of the same height (150 cm) vary from this now that we are including weight in our prediction as well.

Example - Prediction Intervals

First, here is a 95% prediction interval for the FVC of individuals with height 150 cm and weight 41.5 kg:

```
       fit      lwr      upr
1 2.765035 2.16325 3.366821
```

Individuals in the FVC data set with height around 150 cm have weight approximately 41.5 kg on average, so

- the prediction interval centred in the same place as the prediction interval using only height, but
- it is slightly more precise, because we have more confidence in our prediction due to our increased ability to explain variability in FVC with the additional information about weight.

But what if we made a prediction for the FVC of an individual with a different weight?

Example - Prediction Intervals

In the FVC data set, there are individuals with height of 150cm and a weight above 50 kg, so lets calculate a 95% prediction interval for the FVC of individuals with height 150 cm and weight 55 kg is:

```
       fit      lwr    upr
1 2.965794 2.351189 3.5804
```

Notice the prediction interval is shifted to higher FVC, because the slope coefficient b_2 is positive so we predict individuals with higher weight will have higher FVC as well!

Example - Confidence Intervals

We can also predict the population mean FVC of individuals in a similar way by calculating 95% confidence interval for the FVC of individuals with height 150 cm and weight 41.5 kg:

```
       fit      lwr      upr
1 2.765035 2.709035 2.821036
```

and for individuals height 150 cm and weight 55 kg is:

```
       fit      lwr      upr
1 2.965794 2.828935 3.102653
```

Comment on the similarities and differences between the 95% prediction intervals and the 95% confidence intervals, are they what you would expect?

Example - Solution

The point estimates are the same, but the width of the confidence intervals is much less than that of the prediction intervals.

This is exactly what we'd expect, since both prediction and confidence intervals are centred on the point estimate, but we can be more certain about the mean FVC for a population than we can be about the FVC of an individual!

More Then Two Numeric Predictors

We won't go into including more than two numeric predictors, but hopefully you can see how the process of adding a second numeric predictor could be repeated to add a third, fourth, and ultimately as many as you like.

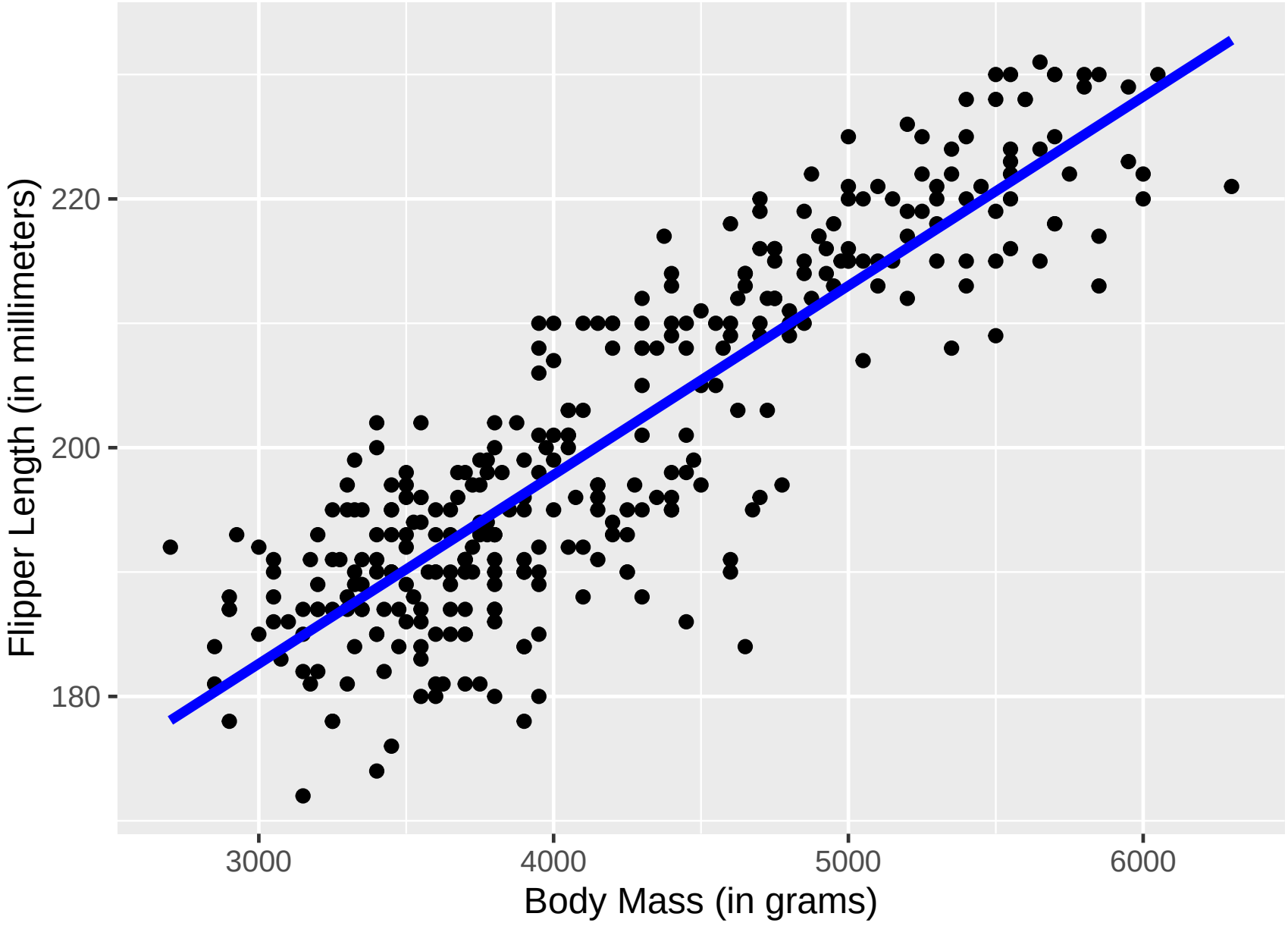
Categorical Predictors

Including Categorical Variables in Regression

So we can add more numeric variables to a regression model, but what about adding categorical variables as predictors?

Remember the palmerpenguins data? It was data on penguins in which we had a number of variables recorded from three species of penguin including body mass and flipper length. We could fit a regression line to predict flipper length based on body mass, both numeric variables, using the methods we introduced in the previous topic:

Example - Penguins



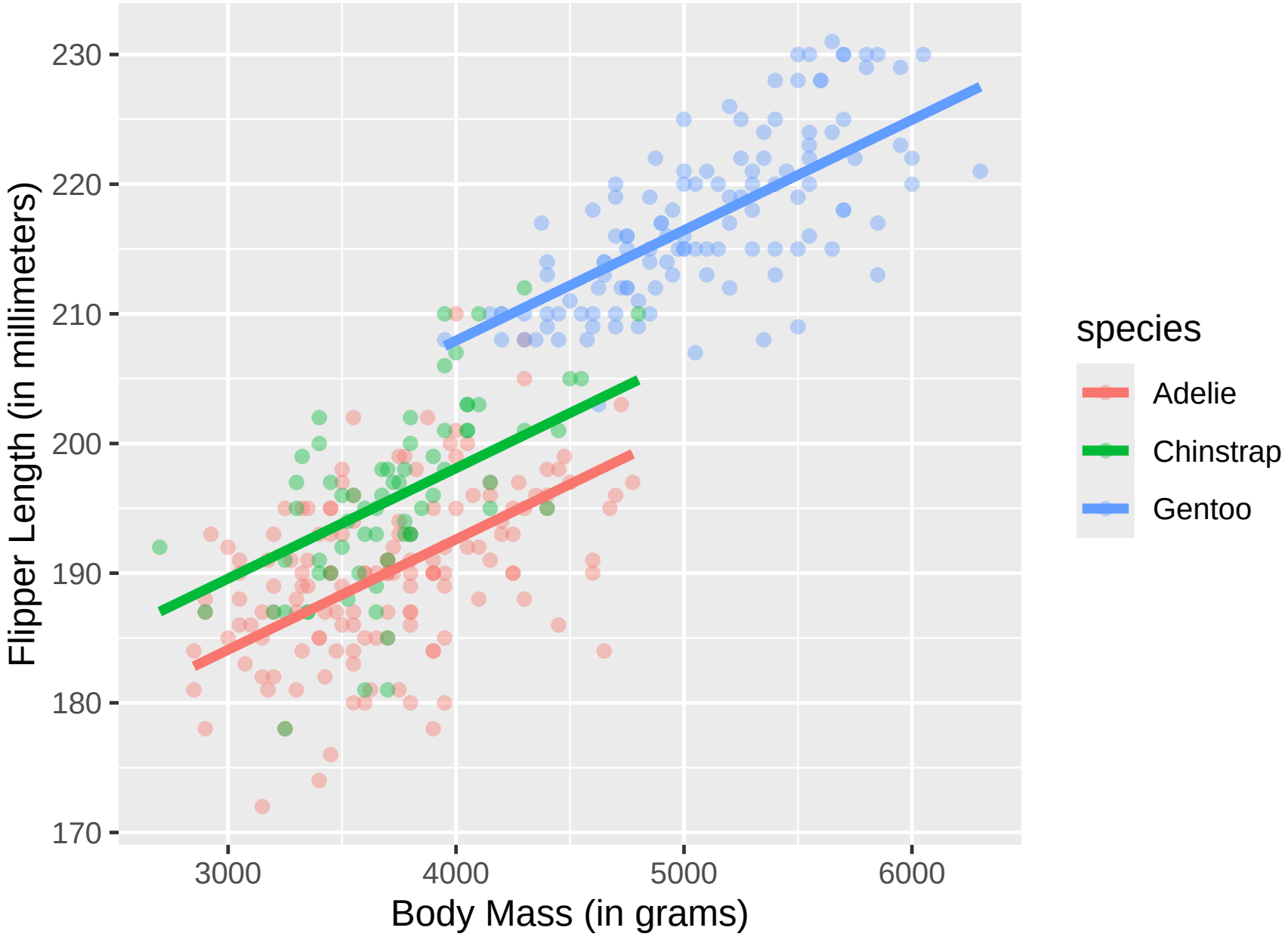
Example - Penguins

But there are three different species of penguins, which might have different associations between body mass and flipper length! So we want to fit three different lines, one for each species.

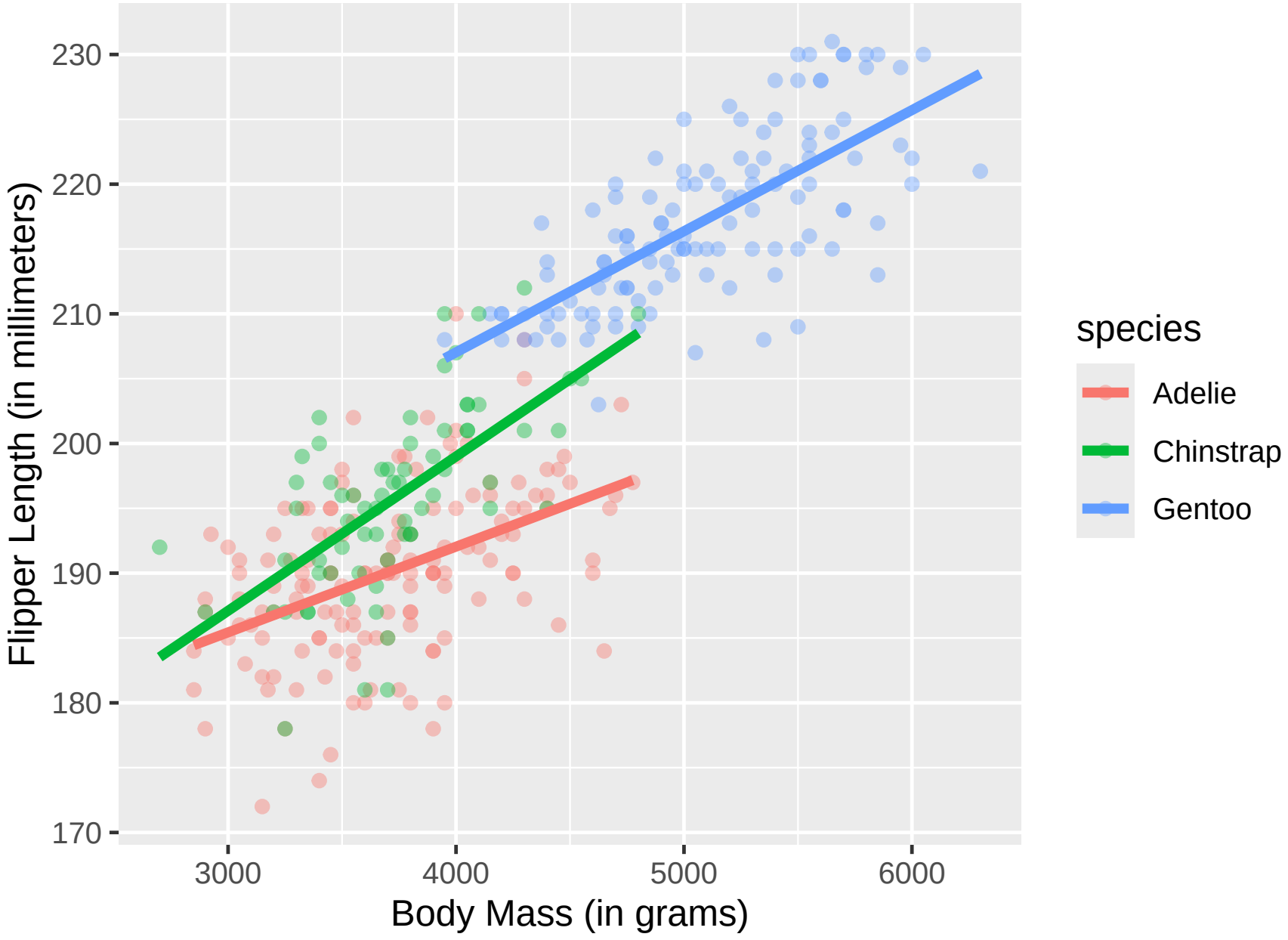
There are essentially two ways to do this, which differ depending on if:

- we assume that the slope of the three lines are all the same or if
- we allow the three lines to have different slopes.

Example - Penguins

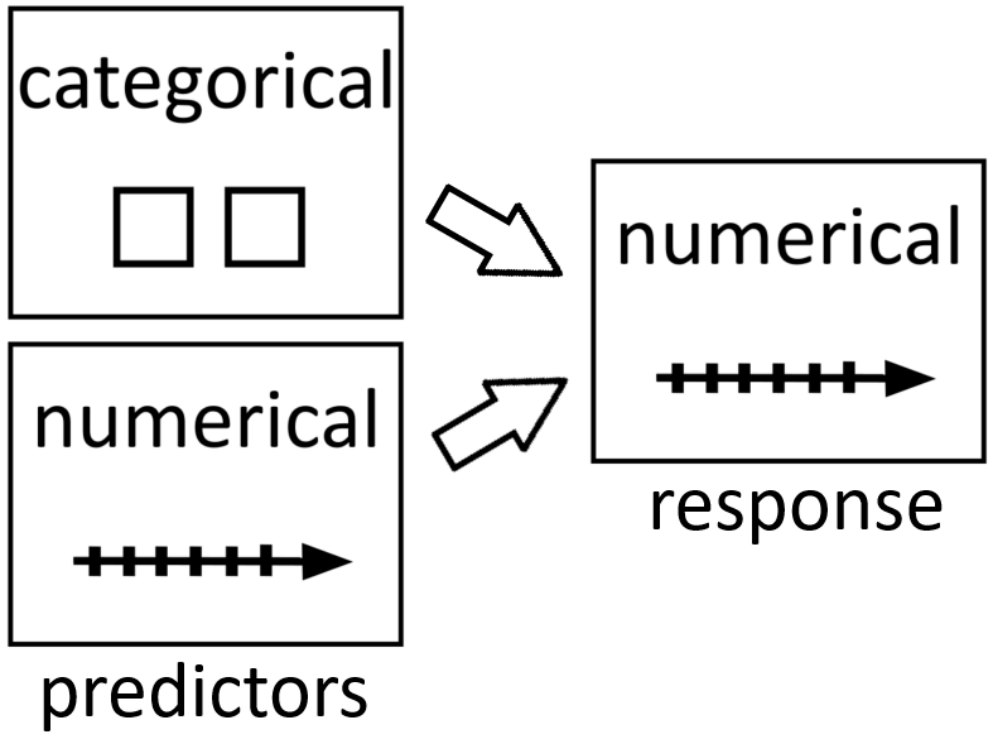


Example - Penguins



One Numeric and One Categorical Predictor

Either way, we are adding a categorical variable — species — as a predictor to our 'simple' linear regression model with one numeric predictor



No Interaction Term

If we assume that the slope of each of the lines is the same this is called the **no interaction** model. We'll come back to why this is when we discuss the case where we allow the lines to have different slopes.

This corresponds to adding term(s) to the equation for one numeric predictor, much like we did for two numeric predictors. However, now we are adding a categorical variable — species in the case of the penguins data. So how do you multiply a category by a number?

Coding Categories with Indicator Variables

We do this by coding the categories with **indicator variables**. These are **discrete numerical** variables that only take two values — 0 or 1.

To code a categorical variable with two categories:

- 0 would correspond to one category, and
- 1 would correspond to the other.

Coding Categories with Indicator Variables

To code more than two categories, say we have c categories, we need $c - 1$ indicator variables. The default way to code categories in R which is what we will use, designates an indicator variable for each category *except the first category*, and then either:

- all indicator variables are 0 indicating the first category, or
- one indictor variable is 1 indicating the category corresponding to that indicator.

Coding Penguin Species

In the penguins data there are three species, which in R will by default be listed in alphabetical order:

- Adelie,
- Chinstrap, and
- Gentoo.

So we create two indicator variables, let's call them x_2 and x_3, corresponding to Chinstrap and Gentoo respectively. Then our regression equation becomes

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

where x_1 is body mass, and y is flipper length.

Coding Penguin Species

The one equation

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

then includes equations for all three lines, one for each species:

- Adelie: $x_2 = 0$ and $x_3 = 0$, so the equation becomes $y = b_0 + b_1 x_1$.
- Chinstrap: $x_2 = 1$ and $x_3 = 0$, so the equation becomes $y = b_0 + b_1 x_1 + b_2 = (b_0 + b_2) + b_1 x_1$, and
- Gentoo: $x_2 = 0$ and $x_3 = 1$, so the equation becomes $y = b_0 + b_1 x_1 + b_3 = (b_0 + b_3) + b_1 x_1$,

which all have the same slope, b_1, but different intercepts b_0 for Adelie, $b_0 + b_2$ for Chinstrap, and $b_0 + b_3$ for Gentoo.

Example - Penguins without Species

So how does this work in practice? Let's start by fitting the 'simple' linear regression model without species:

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

Residuals:
    Min      1Q  Median      3Q     Max
-23.698  -4.983   1.056   5.101  13.933

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.370e+02  1.999e+00   68.56   <2e-16 ***
body_mass_g 1.520e-02  4.667e-04   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.847 on 331 degrees of freedom
Multiple R-squared:  0.7621,    Adjusted R-squared:  0.7614
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

Example - Penguins without Species

So we have approximate values $b_0 = 137$, $b_1 = 0.0152$, and an R-squared of 76%. Our fitted regression line is

$$\text{length} = 137 + 0.0152 \times \text{mass}$$

and explains approximately 76% of the variation in flipper length using body mass, without taking into account species (combining all species together).

So we would predict that if the body mass increases by 1 kg, the mean flipper length would increase by 15.2 mm.

What happens when we add species?

Example - Penguins with Species (no interaction)

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

Residuals:
     Min       1Q   Median       3Q      Max
-14.5418  -3.1804   0.0983   3.3295  17.3954

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.585e+02  2.435e+00  65.119  < 2e-16 ***
body_mass_g      8.515e-03  6.457e-04  13.186  < 2e-16 ***
speciesChinstrap 5.492e+00  7.938e-01   6.918  2.4e-11 ***
speciesGentoo    1.533e+01  1.117e+00  13.727  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.405 on 329 degrees of freedom
Multiple R-squared:  0.8526,    Adjusted R-squared:  0.8513
F-statistic: 634.4 on 3 and 329 DF,  p-value: < 2.2e-16
```

Example - Penguins with Species (no interaction)

The R-squared of 76% without species increases to 85% and we have approximate values $b_0 = 158.5$, $b_1 = 0.0085$, $b_2 = 5.5$, and $b_3 = 15.3$ so our fitted regression equation is

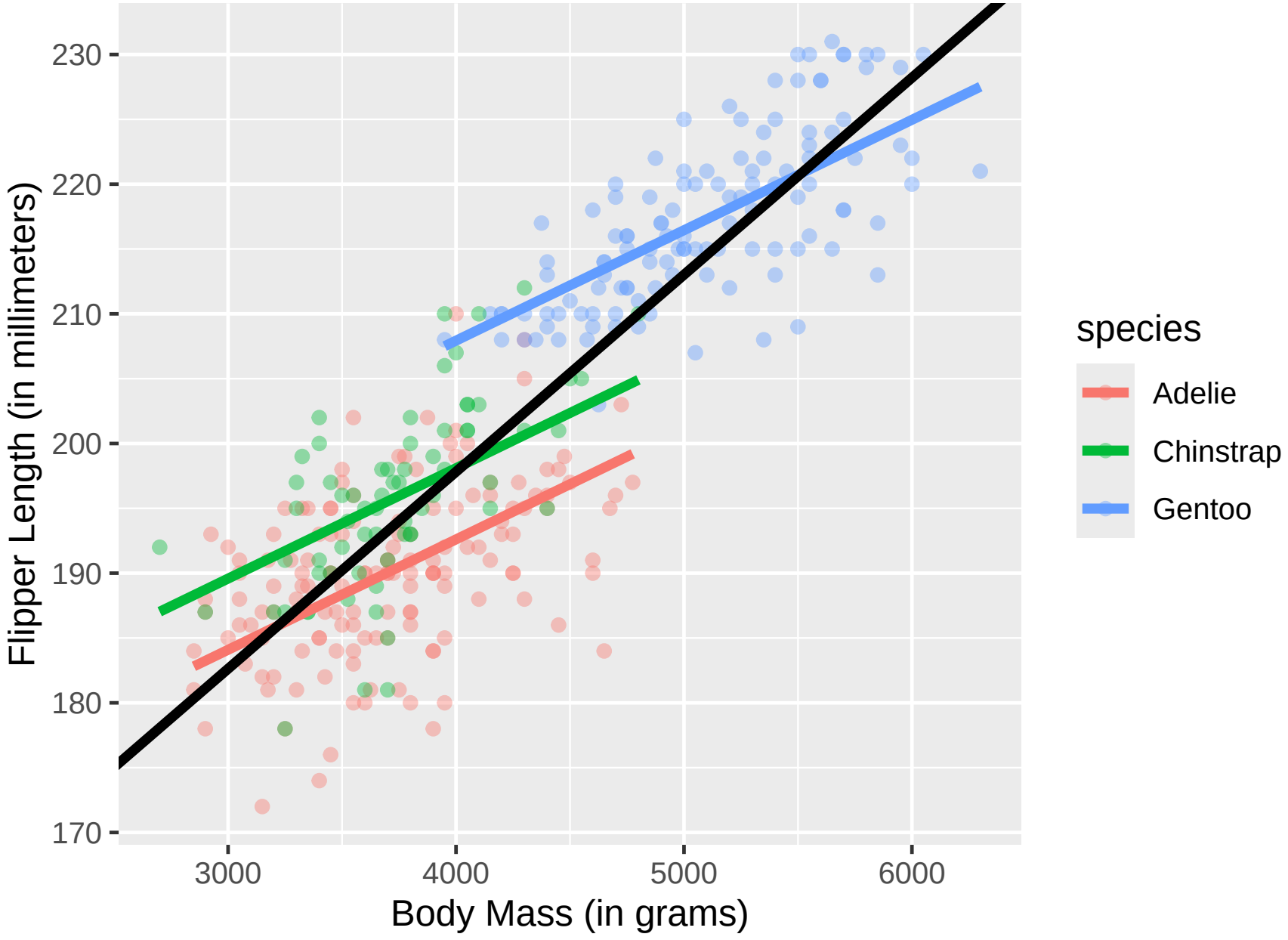
$$y = 158.5 + 0.0085x_1 + 5.5x_2 + 15.3x_3$$

and equations for each species would be:

Adelie	length $= 158.5 + 0.0085 \times$ mass
Chinstrap	length $= (158.5 + 5.5) + 0.0085 \times$ mass
	$= 173.9 + 0.0085 \times$ mass
Gentoo	length $= (158.5 + 15.3) + 0.0085 \times$ mass
	$= 173.9 + 0.0085 \times$ mass

Notice that the slope is much less than when species was not included, and now those changes in flipper length are instead explained by species.

Example - Penguins with Species (no interaction)



Interaction Terms

In order to fit lines with different slopes for each species, we need to include interaction terms in our model.

So what is an interaction term?

An interaction term is when we include the multiplication of two variables in our model, and allows us to model interactions between predictors — in the penguins data example, the interaction between body mass (the numerical predictor) and species (the categorical predictor).

Example - Penguins - Interaction Model

Since we've coded our categorical predictor species with two indicator variables x_2 and x_3 for the Chinstrap and Gentoo penguins respectively, we need to add an interaction term for each of these with the x_1 body mass numeric predictor. Doing so gives us the regression equation

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 + b_5x_1x_3$$

and similarly to the non-interaction model, this one regression equation will include equations for all three lines, one for each species.

Example - Penguins (with interactions)

- Adelie: $x_2 = 0$ and $x_3 = 0$, so the equation becomes

$$y = b_0 + b_1 x_1,$$

- Chinstrap: $x_2 = 1$ and $x_3 = 0$, so the equation becomes

$$y = b_0 + b_1 x_1 + b_2 + b_4 x_1 = (b_0 + b_2) + (b_1 + b_4) x_1,$$

- Gentoo: $x_2 = 0$ and $x_3 = 1$, so the equation becomes

$$y = b_0 + b_1 x_1 + b_3 + b_5 x_1 = (b_0 + b_3) + (b_1 + b_5) x_1.$$

which allows the three lines to have different slopes as well as intercepts by adding b_4 and b_5 in comparison to the no interaction model. Lets fit it and see!

Example - Penguins (with interactions)

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

Residuals:
     Min       1Q   Median       3Q      Max
-14.4296  -3.2343   0.1668   3.2517  17.9549

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.656e+02  3.619e+00  45.756  < 2e-16 ***
body_mass_g                    6.611e-03  9.692e-04   6.820  4.4e-11 ***
speciesChinstrap              -1.422e+01  7.339e+00  -1.938  0.05350 .
speciesGentoo                  4.064e+00  6.195e+00   0.656  0.51227
body_mass_g:speciesChinstrap   5.295e-03  1.958e-03   2.704  0.00721 **
body_mass_g:speciesGentoo      2.730e-03  1.380e-03   1.978  0.04872 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.352 on 327 degrees of freedom
Multiple R-squared:  0.8564,    Adjusted R-squared:  0.8542
F-statistic: 389.9 on 5 and 327 DF,  p-value: < 2.2e-16
```

Example - Penguins (with interactions)

The R-squared of 76% without species increased to 85% with species but without interactions, and now to 86% with interactions.

We have approximate values

- $b_0 = 165.6$,
- $b_1 = 0.0066$,
- $b_2 = -14.2$,
- $b_3 = 4.1$,
- $b_4 = 0.0053$, and
- $b_5 = 0.0027$.

Example - Penguins (with interactions)

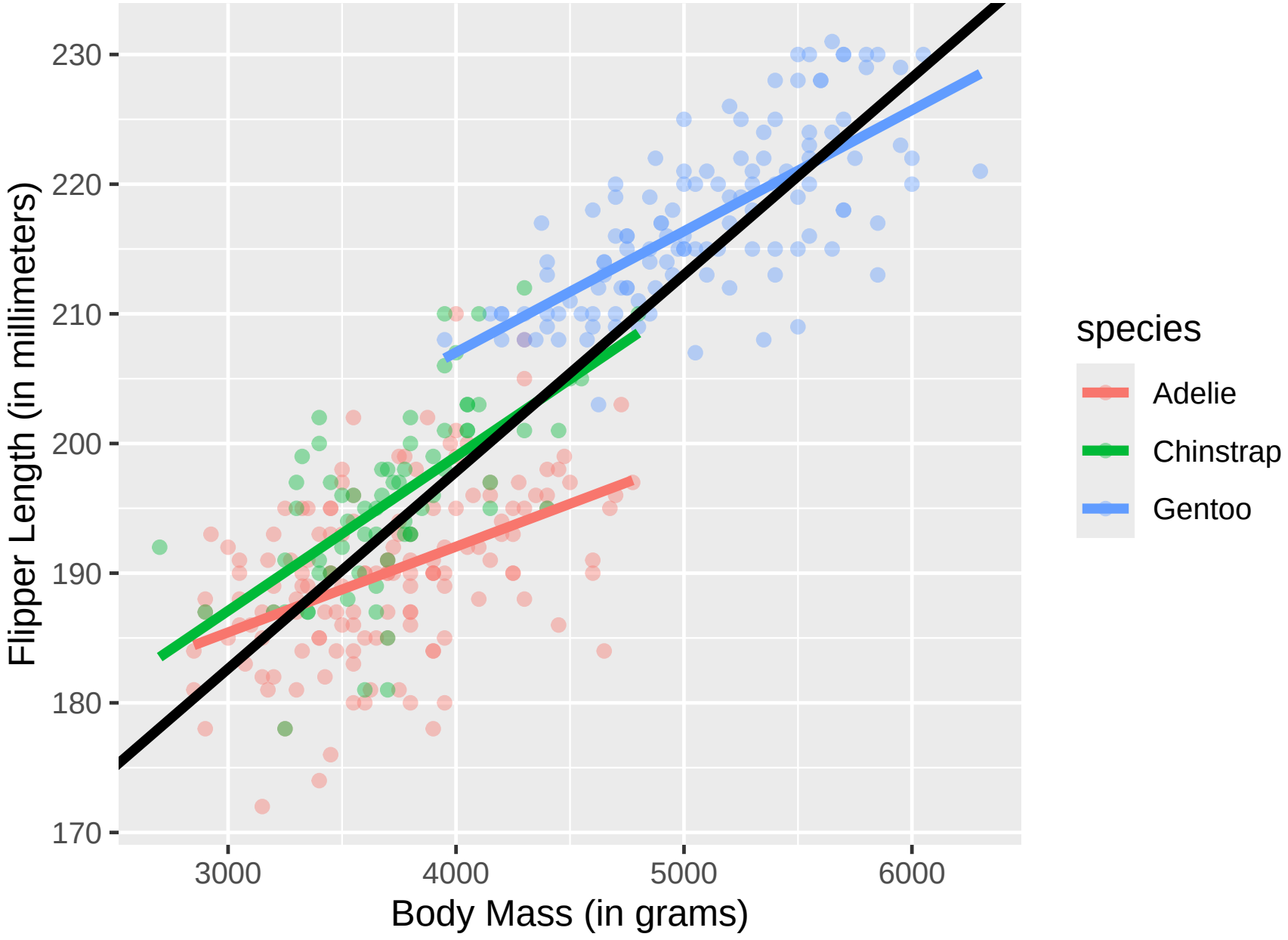
So our fitted regression equation is

$$y = 165.60.0066x_1 + -14.2x_2 + 4.1x_3 + 0.0053x_1x_2 + 0.0027x_1x_3$$

and equations for each species would be:

Adelie	$\text{length} = 165.6 + 0.0066 \times \text{mass}$
Chinstrap	$\text{length} = (165.6 - 14.2) + (0.0066 + 0.0053) \times \text{mass}$
	$= 151.4 + 0.0138 \times \text{mass}$
Gentoo	$\text{length} = (165.6 + 4.1) + (0.0066 + 0.0027) \times \text{mass}$
	$= 169.7 + 0.0112 \times \text{mass}$

Example - Penguins (with interactions)



Assumptions, Inference and Prediction

We can:

- Check assumptions,
- Make inferences about the parameters, and
- Make predictions

using these models much like we did with two numerical predictors.

To avoid being repetitive however, we'll leave going through checking assumptions, making inferences and predictions for these models for you to try in the practical.

Why the Interaction Model?

Fitting three seperate models with only one numeric predictor to each of the three different species seperately would result in the same fitted lines as the interaction model we just showed.

However, doing all this in one model easily facilitates some things that fitting three seperate models doesn't:

- An overall quantification of prediction/ explanatory power (the R-square value) accross all species together,
- Direct comparison between the values for the different species — fitting seperate models to each species wouldn't easily allow this.

But What Does it all Mean?!

The three species of penguins are clearly different. When we explored these data in topic 2: 'Summarising and Visualising', we could see from the scatterplot and summary statistics that Gentoo penguins are larger and have longer flippers overall than the other two species. So what does this model allow us to see that we couldn't already see? Some examples of interpretations this model allows us to make that we couldn't have made otherwise:

- We would expect a Gentoo penguin to have longer flippers than another penguin of one of the other two species, *even if they had the same body mass (say, 4 kg)*.

But What Does it all Mean?!

- The relationship between body mass and flipper length is different between the three species of penguins, for example in particular the Adelie penguins of the three species seem to increase flipper length least for a given increase in body mass.
- Although the Adelie and Chinstrap penguins tend to have similar body masses and flipper lengths overall, it seems they have a different association between these two numerical variables, which is shown in particular by the interaction term between the Chinstrap indictor variable x_2 and body mass .

Which Model Should we Use?

When analysing data like this, there isn't a 'correct' model to use, instead it is important to understand what assumptions we are making when we fit each model so that we know how to interpret the results appropriately and what disclaimers to include.

How Not to Make a Conlcusion

What is wrong with the following conclusion?

> *We can see from the no-interaction model that the rate at which flipper length increases in relation to body mass (the slopes of the three lines) seems to be the same for all three species of penguins.*

Summary and Next Week

Summary

In a regression model, we always have a numeric response variable.

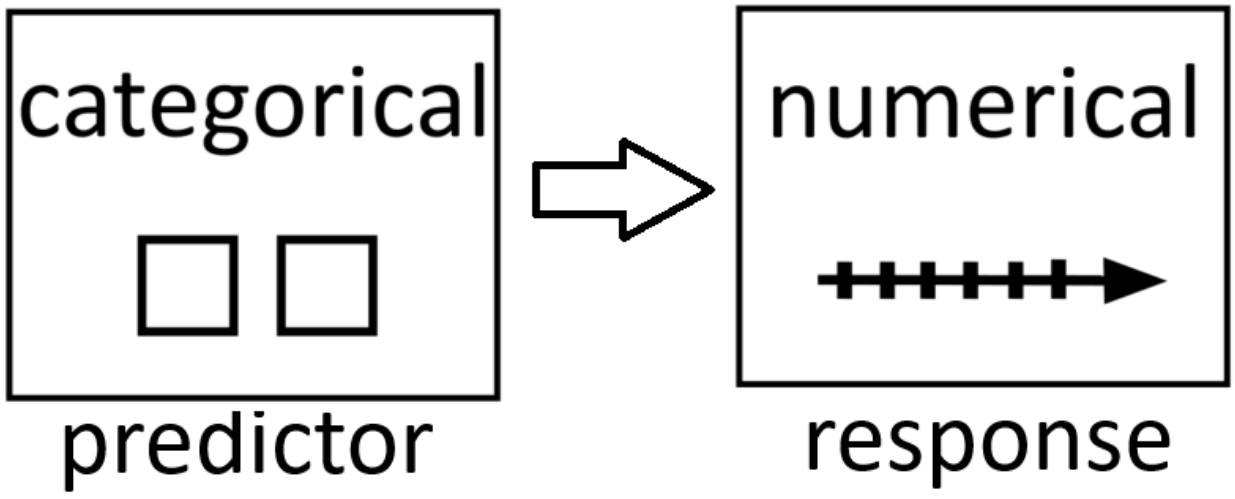
In the previous forum, we introduced regression using a single numeric explanatory variable or predictor, and in this forum we showed how you can add additional numeric or categorical predictors.

But what if you only used categorical explanatory variables and didn't use any numeric predictors at all?

It turns out we've already covered this — the two-sample t-test!

Next Week

Two-sample t-test as regression:



Next Week

In the next forum, we'll introduce:

- how the two-sample t-test can be thought of as regression using one categorical predictor (with two categories which identify the two samples) and
- how this can be extended to compare more than two groups simultaneously in something called 'analysis of variance' (ANOVA), as well as a non-parametric way of doing this called the Kruskal-Wallis test.

The next forum will be the last forum in which we introduce new concepts, and the following (and final) forum we will review and summarise all that we've discussed in this course, and begin preparations for the exam!