Introduction to Linear Models

A worksheet for Data Science. Written by Hannah Wauchope

10th Oct, 2025

Set up

Open the “LinearModels1_Introduction.R” script within this week’s folder in the course repo. Run the ‘set up’ lines that load the packages etc.

```
#We're going to be using the palmerpenguins dataset
library(palmerpenguins)
library(tidyverse)
library(ggeffects)
```

Let’s begin

Exercise 1 – Inspect the data

Take a moment to have a look at the ‘penguins’ dataset that we’ve just loaded with palmerpenguins.

- How many species do we have data for?
- What are the names of the islands where the data was collected?
- What species variables do we have data for?
- How many years of data do we have?

```
#We're going to be using the palmerpenguins dataset
summary(penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

```
#We have data for 3 species, at 3 islands (Biscoe, Dream and Torgersen), we have da
ta on bill lengt and dept, flipper length and body mass. We have data for 2007, 200
8 and 2009
```

Now, to help with interpretations in a minute, I’d like you to create a new variable representing body mass in kilograms, not grams. Call it “body_mass_kg”. Please do this now, and check with a friend/tutor if you’re not sure what the code should be.

```
#If you did it in base r it would look like this:
penguins$body_mass_kg <- penguins$body_mass_g/1000

#Or, if you did it in tidyverse, it would look like this:
penguins <- penguins %>% mutate(body_mass_kg = body_mass_g/1000)
```

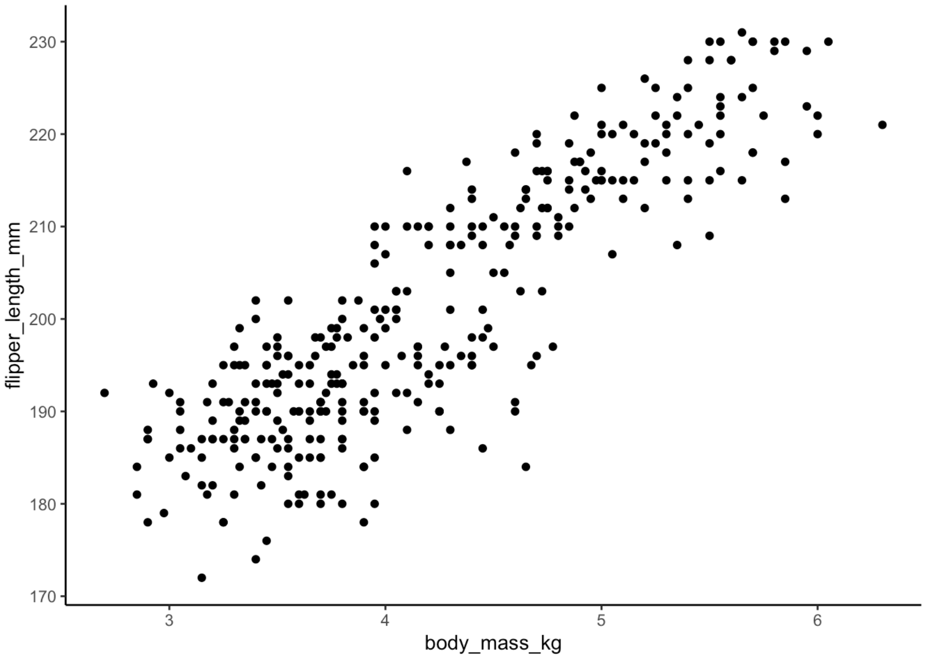
Exercise 2 - Look at relationships

I’m interested to know if heavier penguins (body mass *in kilograms*) have longer flippers. Let’s start by just making a plot comparing these two variables (remember ggplot is your friend! And a hint: geom_point is also your friend)

First, *hand draw* here what you would want that plot to look like, then code it after

```
ggplot(data=penguins, aes(x=body_mass_kg, y=flipper_length_mm))+
  geom_point()+
  theme_classic()
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale ra
nge
## (`geom_point()`).
```



Build A Simple Model, with One Continuous Variable

Exercise 3 - Model relationships

Looks to be a pretty strong relationship. Let's model it! (Why? It could be to see how confident we are in what we see, or it could be because we want to make estimates - "how much longer are a penguin's flippers, on average, for every kilogram heavier they are?")

Build a simple linear model comparing our variables. Here's some base code for a simple linear model, substitute in the right variable and data names for your code (remember to use the kg body mass variable!):

```
Model1 <- lm(response ~ predictor, data=MyData)
```

```
Model1 <- lm(flipper_length_mm ~ body_mass_kg, data=penguins)
```

Check Assumptions

Exercise 4 - Remember assumptions

Now then. I know the temptation. You wanna get straight to seeing the p value. That is NOT the most interesting thing about the model!

Before we can even get to seeing what the model says, we need to see how it fits the assumptions. Why? If it doesn't fit them well, then what it tells us isn't very trustworthy and the model isn't much use. We shouldn't even look at the results unless we're happy they're going to tell us something sensible.

To that end, write here what the five assumptions are that we need to check for. Discuss with your partner what each of these actually mean, and why we should care 1. 2. 3. 4. 5.

One of these assumptions is not something we can check with code, we just need to consider it based on our knowledge of the data. What is it, and do you think we've violated it?

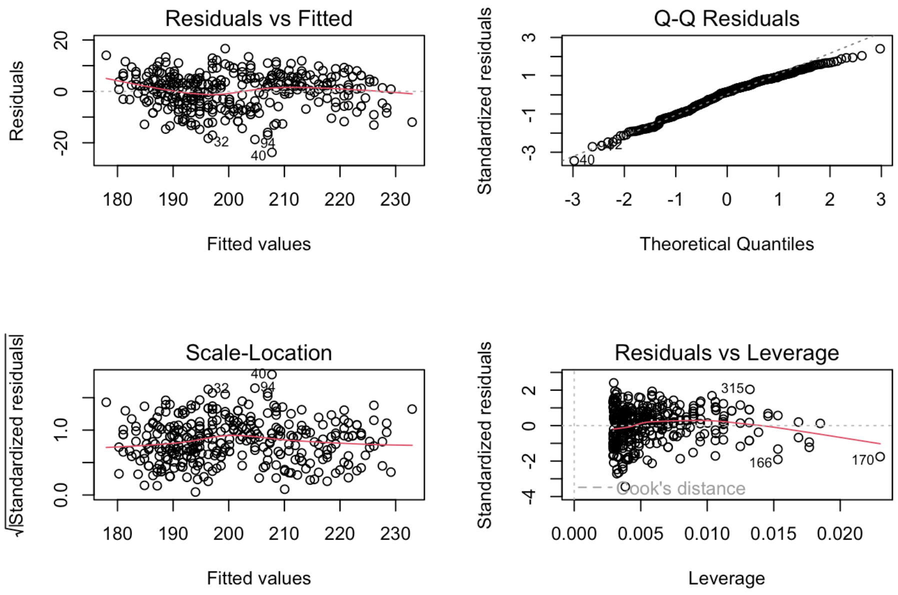
```
#The assumptions we need to check for are:
#1. Independence
#2. Linearity
#3. Normality
#4. Constant Variance
#5. Outliers

#We can't check for independence using code, we just have to consider it based on our knowledge of how the data was collected
```

Exercise 5 - Test for assumptions

Okay, let's see how the model does on violating the other four assumptions. Run the following code:

```
par(mfrow=c(2,2)) # Change the panel layout to 2 x 2 so you can look at all the plots at once (it's less annoying)
plot(Model1)
```



```
par(mfrow=c(1,1)) # Change the plot setting back to 1 x 1
```

Write here what assumption each of these plots checks for:

- Residuals vs Fitted:
- Q-Q Residuals:
- Scale-Location:
- Residuals vs Leverage:

Now, inspect your plots, and discuss with your partner/tutors. How do you think we're doing on these assumptions, have we violated any?

```
#See here for great explanations! https://library.virginia.edu/data/articles/diagnostic-plots Also Frank Rodenburg's video (in flipped classroom learning materials)

#Residuals vs Fitted - LINEARITY: You should see a clear trend from left to right to be worried about violating this assumption, don't stress too much about the line - it's very reactive
#QQplot - NORMALITY: Normality if residuals are more or less normally distributed, they'll more or less follow the line. Often you see little upticks at the top and bottom, this is okay. This can be checked with the Shapiro Wilk test, but we're focussing on visual inspection today.
#Scale-Location - CONSTANT VARIANCE: Again, don't stress too much about the line if it's wiggly. But if it's clearly and consistently going up or down, that means variance is clearly going up or down with our predictor. In this cae fine.
#Residuals vs Leverage - OUTLIERS: None looking too crazy here. When things are bad you see red dotted lines, and one or two points on the other side of them to the rest - These are major outliers. That doesn't necessarily mean you should just remove them!! If you think they're accurate (and not a mismeasurement or someting), then they're an important part of your study system. But know what they'll be having a disproporionate influence on the conclusions of your model - so the model isn't so much representative of what's going on overall in your data.
```

Interpret Model Output

Exercise 6 - Inspect model output

You should have concluded that there's one assumption (out of the five) that we've violated. For now we're going to ignore that for the purposes of teaching, and come back to it in a minute. If you're not sure which assumption it was, go back and check with a tutor.

Okay, now let's look at the model output!

Run `summary(Model1)`

Lots to see here. Remember what I've said in the lecture - this output was not designed with the naive viewer in mind. First of all, we'll focus on the R-squared (near the bottom). You get given two, just focus on the "Adjusted R-squared". Write what it is here, and what it means:

Pick the right word: Body mass of penguins predicts quite a (lot/little) the variation in their flipper length. Fill in the gap: ___% of the variance in penguin flipper length can be explained by their body mass.

What if the R-squared was 0.01? What would that mean about what we think of the relationship between flipper length and body mass? Would we be confident about predicting a penguin's flipper length if we only knew its mass?

```
#R-squared is 0.76. This means that 76% of the variance in flipper length is explai
ned by body mass. The simple way of saying that is that if we get a bunch of pengui
ns, they'll all have different flipper lengths, varying about some average across a
ll of them. But we can explain 76% of how much they vary about that average by just
knowing how heavy each penguin is.

#In ecology, that number is really high! Normally data is so noisy (with some many
complex factors) that variables don't have such a strong relationship. If you were
a physicist though you'd probably be horrified at how low this is - they expect the
y data to behave much more than we do.

#If R-squared was 0.01 then body mass would explain very very little of the varianc
e in flipper length. Even if 'body mass' ended up being significant in the model, w
e would say that it's not an important factor related to flipper length.
```

Exercise 7 - Looking at the relationship

Nope, we're still not gonna look at the P value. Let's look at the 'Coefficients' table. There are 4 columns ('Estimate' 'Std. Error', 't value', and 'Pr(>|t|)'. We're going to have a look at the Estimate column.

Before we do, I'd like you to draw an empty graph here (x axis with 'Body Mass (kg)', y axis with 'Flipper Length (mm)'). Mark the x axis in increments starting at 0 and going up to 7. Add increments to the y axis, starting at 0, and going up in jumps of 50 to 250.

Now, we're going to draw the trend line the model has created (roughly, don't stress about being too exact). Where should the trend line start on the y-intercept? (Hint, the clue is in the word 'intercept').

Remember that when we have a linear predictor, it's 'Estimate' says "for every increase of 1 in the predictor, the response is going to change by ____ amount". Fill in what that '___' should be, based on the model output. Now have a go at drawing the trend line on your graph.

```
#The estimate for body mass is 15.28, meaning for every rise in 1kg of penguin body
mass, flipper length increases by 15.28mm on average. In this case best to report t
he change as 15mm on average - as .28 of a mm is not a specifity we would tend to m
easure!
```

Exercise 8 - Standard Error

So that's the main trend line drawn. But what about the error around that estimate? Now, you can see that the model gives "Std. Error", but it's not best practice to use that - we tend to favour 95% confidence intervals (i.e. 95% of the time we'd expect our estimate to fall somewhere in this range if we kept running the model). To get this, run `confint(Model1)` . What you get is the upper and lower confidence bounds for our intercept and body mass variables. Use these values to shade some confidence interval bars onto your graph.

IMPORTANT NOTE - remember that these confidence intervals represent our confidence in the *average estimate*, not how much variance there is in the data. Especially in large datasets you can have huuuuuuge variance in the data, but a highly confident average estimate. What does this mean? If we have very high variance, but also high confidence in the estimate, that means we're confident in the direction of the *average* relationship between the variables, but that if you tried to predict y based on x for any one individual, you wouldn't be confident at all in your prediction. (Classic example: we know there is an average relationship, across thousands of people, with amount of bacon consumption and life expectancy. But if you tried to predict how long someone is going to live based on how much bacon they eat, there's going be a lot of uncertainty in that prediction!)

Now, let's check whether you got this all right using the `ggpredict` function (from the ggeffects package that we loaded earlier). I *love* this function.

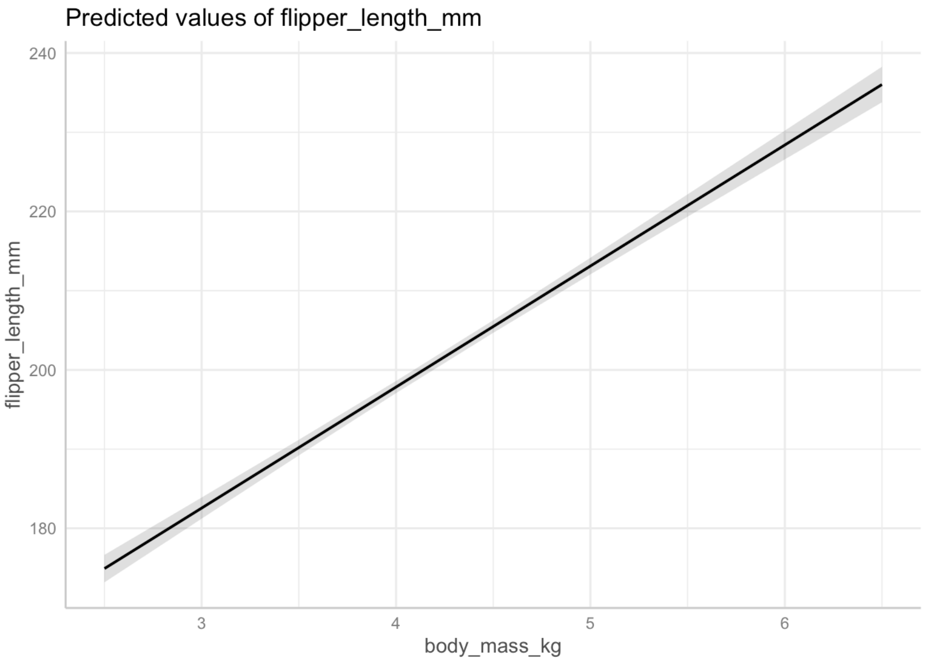
Run the following:

```
PredictedFlippers <- ggpredict(Model1, c("body_mass_kg"))
PredictedFlippers
```

x	predicted	std.error	conf.low	conf.high	group
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
2.5	174.9193	0.8780015	173.1923	176.6463	1
3.0	182.5573	0.6741642	181.2312	183.8834	1
3.5	190.1953	0.4970674	189.2175	191.1730	1
4.0	197.8332	0.3855159	197.0749	198.5915	1
4.5	205.4712	0.3989197	204.6865	206.2558	1
5.0	213.1091	0.5278440	212.0709	214.1474	1
5.5	220.7471	0.7120882	219.3464	222.1477	1
6.0	228.3851	0.9189601	226.5775	230.1926	1
6.5	236.0230	1.1361663	233.7882	238.2578	1

9 rows

```
plot(PredictedFlippers)
```



Exercise 9 - P Values

Finally, let's take a look at the p values (the bit saying Pr(>|t|)). You can see the 'Signif. codes' below the main table, these just tell you what the asterisks mean. Remember that <0.05 is just a convention, it doesn't really mean anything! It just means that the model thinks if we were to go out and collect the data over and over again, 95% of the time we think we would still find the same relationship (yes that is an odd way of thinking about it, that's why some people prefer Bayesian Statistics, where we just think about how likely it is that we've got the results we do, not this weird thinking of repeated data collection). We especially can't rely on p values in data science, because large datasets almost always get significant results, and sometimes they're pretty meaningless - what's more interesting is the strength of the relationship and the variance around the average.

Nevertheless, is body mass a 'significant' predictor of flipper length, at a conventional p<0.05 cutoff?

```
#Yes, at P <0.0001
```

Exercise 10 - Concluding this section

Please summarise your findings here. What have you found about the relationship between body mass and flipper length? Is it positive or negative? How strong is the relationship? How confident are you in the relationship? Are you satisfied your model is appropriate?

```
#There is a strong positive relationship between body mass and flipper length, with
flipper length increasing by 15mm on average for every 1kg increase in body mass (9
5% CI: 14-16mm). Body mass explains a high proportion of the variance in flipper le
ngth (76%). The model meets all assumptions*, meaning we can be confident in these
results.

#*(well, except Independence, but we're pretending here)
```

Build a Simple Model, with One Categorical Variable

Exercise 11 - Categorical Variables

Rightio. Hopefully earlier you identified that the one assumption we violated was independence: our data was collected on three islands, from three species. Data on Adelie penguins on Biscoe Island are going to be more similar to each other than, for instance, data on Chinstrap penguins from Dream Island.

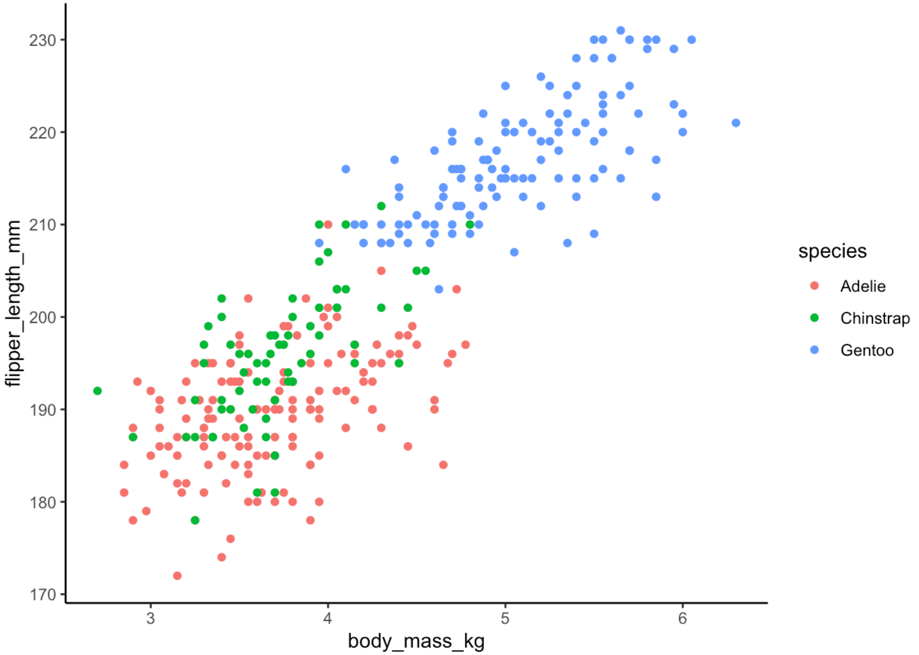
We need to address this, and also it might be interesting to know more about whether flipper length varies by species, not just body mass! Let's make a model that considers the how both species *and* body mass relate to flipper length

[Note that technically to address the independence assumption we should consider island also, but for simplicities sake let's pretend all the data came from one island].

First of all, let's plot our data. Go and retrieve your ggplot code from the start of this exercise. Within the aes, add a line saying `colour=species` .

```
ggplot(data=penguins, aes(x=body_mass_kg, y=flipper_length_mm, colour=species))+
  geom_point()+
  theme_classic()
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale ra
nge
## (`geom_point()`).
```



Interesting - does it look like species affects average flipper length?

Exercise 12 - Categorical Variables Model

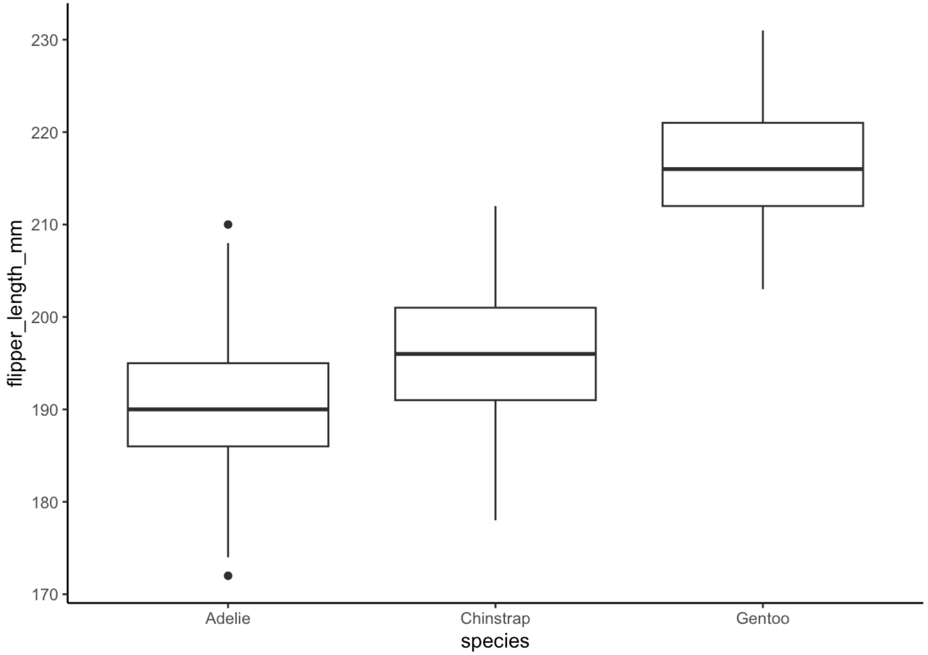
We're going to build up in stages. First, let's look at ONLY the relationship between species and flipper length.

Can you build a boxplot displaying that data - boxplots are best for categorical variables (hint: take the code from the exercise above, make 'x' 'species', and change to geom_boxplot()).

Now can you make a simple linear model comparing these two variables?

```
ggplot(data=penguins, aes(x=species, y=flipper_length_mm))+
  geom_boxplot()+
  theme_classic()
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```
Model2 <- lm(flipper_length_mm ~ species, data=penguins)
```

Hold your horses before looking at the model output. Plot the assumptions, and take a beat. List them here, and check each associated plot to be sure you're happy with how things are looking. You'll notice your points are *very* clustered in these plots - that's cos we're using categorical variables. What you're looking for (in Residuals vs Fitted, Scale-Location, and Residuals vs Leverage) is that the points for each category are roughly evenly distributed above and below the dashed line.

1.
2.
3.
4.
5.

Exercise 13 - Interpreting Categorical Variables Model

Now run `summary(Model2)`. (Adjusted) R-squared first: tell me - how well does penguin species explain average flipper length?

And now let's look at the 'Coefficients' table. What's weird about this table?

That's right - Adelie penguins are missing. This is because, with categorical variables, the Intercept *represents* one of the variables, and all other output is given relative to that. Based on that information, can you tell me:

- What is the average flipper length for Adelie penguins?
- What are the 95% confidence intervals for this estimate? (remember: `confint(Model2)`)
- Is this estimate significantly different from zero?

```
#Adjust R-squared is 0.78. Penguin species explains 78% of the variance in flipper
length, even more than body mass!

#What's weird about the table is that 'Adelie' is missing

#Adelie = 189.95
#Intervals = 188.89 - 191.01
#Yes, p < 0.0001
```

And now, the estimates we get for the other categorical variables are *relative* to the estimate for Adelies. Time to do some maths!

- What is the average flipper length for Chinstrap penguins? (Hint: you need to add two numbers together)
- Is this significantly different to Adelies?
- What is the average flipper length for Gentoo penguins?
- Is this significantly different to Adelies?

```
#Chinstrap = 189.95 + 5.86 = 195.81
#Yes, p < 0.0001
#Gentoo = 189.95 + 27.23 = 217.81
#Yes, p < 0.0001
```

Now run `ggpredict(Model2, c("species"))` , and then plot it. Did you get things right?

Build a Model with Multiple Predictors (Continuous and Categorical)

Exercise 14 - Bring it home

You're doing so well. The last thing we're going to do is bring all our learnings together. The beauty of linear models is that you don't have to model one thing at a time. I want to know how species AND body mass predict a penguin's average flipper length. Can you create that model? (Hint: you'll need a plus sign)

```
Model3 <- lm(flipper_length_mm ~ species + body_mass_kg, data=penguins)
```

Have a check of those assumptions. Are they looking okay?

```
#All look good!
```

Now run `summary(Model3)` . What does the Adjusted R-squared tell us? Which of the three models we've run today has done the best job at explaining variance in flipper length?

```
summary(Model3)
```

```
##
## Call:
## lm(formula = flipper_length_mm ~ species + body_mass_kg, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5455  -3.1845   0.1307   3.3533  17.5313
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      158.8603     2.3866  66.564  < 2e-16 ***
## speciesChinstrap   5.5974     0.7882   7.101 7.33e-12 ***
## speciesGentoo     15.6775     1.0907  14.374  < 2e-16 ***
## body_mass_kg       8.4021     0.6339  13.255  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.395 on 338 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8541, Adjusted R-squared:  0.8528
## F-statistic: 659.4 on 3 and 338 DF,  p-value: < 2.2e-16
```

```
#Adj R squared is 0.85 = this is the best model so far, and explains the most varia
nce - 85%!
```

Now, I want you to draw your output again. Create an empty graph with axes like you did last time.

Now, the key thing to remember for models with multiple predictors is that, for all categorical estimates, the continuous estimates are *held at zero*.

They're what now? This means that where in Exercise 13 '(Intercept)' told us the average flipper length for Adelie penguins, now (Intercept) tells us the average flipper length for Adelies *when body mass equals zero* (I agree, that's not very helpful. A penguin of body mass equals 0 is a non-existent penguin. But there it is, it's how the model output works).

So, mark on the y axis of your graph what the y-intercept is for Adelie penguins (i.e. when x equals 0). Now mark what it is for Chinstraps (hint - you'll need to add two numbers together again). Now mark what it is for Gentoos (get adding!).

Now, you can draw your trend line. Look at what the estimate is for body_mass_kg. As before, that's how much flipper length rises per 1kg rise in body mass, and this estimate is the same regardless of species. Draw out your lines for each species.

```
#Adelie y-intercept is 158.86
#Chinstrap y-intercept is 158.86 + 5.59 = 164.45
#Gentoo y-intercept is 158.86 + 15.68 = 174.54

#All rise by 8.4 per kg of body mass
```

Check your work by running `plot(ggpredict(Model3, c("body_mass_kg", "species")))` (ggpredict is cheeky and doesn't draw the x axis all the way back to zero, so don't be fooled by that).

Finally, I want you to look back on the initial model we ran (with just body mass) and to your plot from Exercise 11. Do you have any thoughts about how this model output compares to it?

```
#This new model makes estimates in spaces we don't have data for, e.g. Adelie and C
hinstraps penguins don't bigger than about 5kg, but we get estimates there anyway.
You've gotta be careful about this kind of thing when you have continuous and categ
orical variables combined in a model as these extrapolations might not make much se
nse in some cases!
```