R Practical 6: General Linear Models

Alice Risely

20/11/2023

Table of Contents

[Learning outcomes for today 1](#_Toc182836873)

[Preliminary tasks 2](#_Toc182836874)

[Background: Univariate and multivariate models 2](#_Toc182836875)

[General Linear Models (GLMs) 3](#_Toc182836876)

[Q1) Examining predictors of tiger weight 3](#_Toc182836877)

[Install packages 3](#_Toc182836878)

[Recap: Import and explore tiger dataset 3](#_Toc182836879)

[Summarize data 4](#_Toc182836880)

[Plot your data 5](#_Toc182836881)

[Fitting univariate models 6](#_Toc182836882)

[Fitting a GLM 8](#_Toc182836883)

[Interpreting a GLM output 9](#_Toc182836884)

[The intercept and reference groups 9](#_Toc182836885)

[Estimates 9](#_Toc182836886)

[Effect sizes - which variable has the strongest effect? 10](#_Toc182836887)

[Standard errors (a measure of confidence in the estimate) 10](#_Toc182836888)

[Model fit 11](#_Toc182836889)

[Model assumptions 12](#_Toc182836890)

[Presenting GLM results as a table 12](#_Toc182836891)

[Presenting model results as a figure 13](#_Toc182836892)

[Writing conclusions based on GLMs 13](#_Toc182836893)

[Q2) Examining the predictors of penguin weight 14](#_Toc182836894)

## Learning outcomes for today

* Learn the difference between univariate and multivariate models
* Learn about the most common multivariate model: the General Linear Model (GLM)
* Recap: Load and explore the captive tiger dataset
* Recap: Conduct univariate models on tiger weight
* Conduct a generate linear model (GLM) on tiger weight
* Learn how to interpret the output of a GLM
* Learn how to present the output of a GLM
* Apply what you have learned by fitting a GLM to model penguin weight

## Preliminary tasks

1. **If you have not completed practicals 1 to 5, you need to do these first.**

2. If you are ready to start Practical 6, **DOWNLOAD the sixth practical worksheet** from Blackboard, do not just view it. To download it, click the “…” next to the file and click ‘Download original file’.

3. **Save this file in an appropriately named folder on your OneDrive** (e.g., “MODULE NAME/WEEK X”) so you can access it at home.

4. If you are using the cloud based version of R, **navigate on your web browser to** [**<https://login.rstudio.cloud/>**](https://login.rstudio.cloud/) andsign in. If prompted, click “Posit Cloud”. If you are using the desktop version of R, open your current project.

5. **You should already have an existing R studio project from the first practical called ‘R workshop’,** click on this to enter your saved workspace.

6. **Start a new R script for this practical.** Once your project has loaded (it takes a few seconds), click ‘File’ then ‘New file’, then ‘R Script’.

7. **Press the Save button and save your script as “Practical\_6.R”.** This will save all your code you write today.

## Background: Univariate and multivariate models

So far, in this course you have largely been using UNIVARIATE models. Univariate models are those that looks at the **relationship between two variables** (e.g, when you are interested in knowing the impact of ONE variable on another variable).

**In contrast, MULTIVARIATE models are those that examine the effect of MULTIPLE variables on your variable of interest.**

All the statistical tests you have learned about so far have applied univariate models. T-tests, and one-way ANOVAs are all examples of test that fit simple, univariate models to your data.

**However, it is very rare in nature that you have such simple data**! In most cases, there are multiple sources of variation in your data, you usually you would want to capture that complexity using a multivariate model. You also may want to statistically control for any confounding variables in your study.

For example, you want to know the effect of diet on captive tiger weight. But weight might also be affected by age and sex, and the effect of these variables might obscure the real relationship between weight and diet.

You could fit multiple univariate models to look at the independent relationship between weight and diet, antibiotics, parasite diversity, age, and sex. But this will give you biased results as these models can’t account for interactions in your data. It is also very clunky - why fit five models when you can just fit one?

A better way to do it would be to *control* for other variables, such as age and sex in your model, in order to get a more accurate estimate for the effect of parasites on weight. Therefore, you need a multivariate model!

## General Linear Models (GLMs)

General linear models are the most common multivariate models used in statistics. This is because they are extremely flexible - they can handle all sorts of data types and your data does not have to be normally distributed for you to use them (although there are other assumptions that we will get to later).

To fit a general linear model in R, you need the notation:

***m1 <- lm(response ~ explanatory1 + explanatory2 + explanatory3, data)***

This will fit a model that quantifies the effect of explanatory variables 1, 2 and 3 on your response variable. You need to save your model. In this case, I have saved it as an object called ‘m1’.

Next, lets use the tiger dataset to look at the impact of diet treatment on tiger weight, using the dataset tiger\_weight.csv.

We will go through some of the basics first just as a recap.

## Q1) Examining predictors of tiger weight

**Goal: you want to know the effect of diet treatment on tiger weight, whilst controlling for all other variables that might influence weight.**

I will provide most of the R code you need to examine the predictors of tiger weight, including fitting and interpretting a GLM. Later, I will give you a new dataset and you will need to fit and interpret a GLM yourself.

## Install packages

First - we will need some new R packages. Lets install and load them. If you have already installed any in previous sessions, you don’t need to install them again.

# install new packages. If you have already installed any of these before, you do not need to install them again.   
  
install.packages("ggpubr") # package for presenting multiple plots  
install.packages("sjPlot") # package for presenting GLMs  
install.packages("broom") # package for tidying up model outputs  
install.packages("palmerpenguins") # package containing an exmaple dataset

# load packages into R environment  
  
library(ggplot2)  
library(ggpubr)  
library(sjPlot)  
library(broom)  
library(palmerpenguins)

## Recap: Import and explore tiger dataset

In previous praticals (prac 3) you analysed the dataset on captive tigers.



**This dataset includes data from a study that looked at the effect on antibiotic treatment and diet treatment on captive tiger weight and parasite load across zoos in the United States**. Parasite load was measured as the number of parasite eggs found in tiger faecal samples. The scientists also collected data on tiger age and sex.

#### Exercise 1

**Import the tiger data (tiger\_data.csv) from this weeks Blackboard folder.**

You should now how have tiger\_data in your working environment. Lets go through the usual process of having a quick check of the dataset to make sure everything looks ok.

head(tiger\_data) # examine first 6 rows of dataset  
  
str(tiger\_data) # what data format are the variables in?  
  
dim(tiger\_data) # What are the dimentions of the dataset? This tells you how many rows (110) and how many columns (7) there are  
  
names(tiger\_data) # what are the column names of the dataset? You will need to know these for your statistical analyses

## Summarize data

Lets summarise our key variables. Remember, to summarise numerical variables you need the function *summary()*, and to summarise categorical variables you need the function *table()*.

# summarise numerical variables using summary() function  
  
summary(tiger\_data$weight\_kg) # summary of tiger weight  
  
summary(tiger\_data$parasite\_load) # summary of parasite load  
  
# summarise categorical variables using table() function  
  
table(tiger\_data$antibiotic\_treatment) # frequency of differnt antibiotic treatments  
  
table(tiger\_data$diet\_treatment) # frequency of different diet treatments  
  
table(tiger\_data$sex) # requency of different tiger sexes

## Plot your data

You want to know if diet treatment is significantly associated with tiger weight. But - weight could be influenced by loads of things! It could be affected by how many parasites a tiger has, it’s age, its sex, whether it had antibiotic treatment. Lets plot all the potential UNIVARIATE relationships with weight.

We will save all the plots as objects and then plot them together using the **ggarrange()** function.

# Plot weight ~ antibiotic treatment  
  
p1<-ggplot(tiger\_data, aes(x = antibiotic\_treatment, y = weight\_kg))+  
 geom\_boxplot()  
  
# Plot weight ~ diet treatment  
  
p2<-ggplot(tiger\_data, aes(x = diet\_treatment, y = weight\_kg))+  
 geom\_boxplot()  
  
# Plot weight ~ sex  
  
p3<-ggplot(tiger\_data, aes(x = sex, y = weight\_kg))+  
 geom\_boxplot()  
  
  
# Plot weight ~ parasite load  
p4<-ggplot(tiger\_data, aes(x = parasite\_load, y = weight\_kg))+  
 geom\_point()+  
 geom\_smooth(method = "lm")  
  
# Plot weight ~ age  
  
p5<-ggplot(tiger\_data, aes(x = age\_years, y = weight\_kg))+  
 geom\_point()+  
 geom\_smooth(method = "lm")

## plot all together  
  
ggarrange(p1, p2, p3, p4, p5) # ggarrange() is a function belonging to the ggpubr package we installed earlier

#### Exercise 2

Examine all the plots above. Which variables do you think have a significant association with weight, and what is the direction of the association for each one?

\*\* Write your answers here \*\*  
  
  
  
---

## Fitting univariate models

Above you have made five plots. That means we could conduct five univariate statistical tests to understand the effect of each variable on weight. Lets fit them seperately.

To choose the appropriate univariate statistical tests, you need to know if your numerical variables are normally disitrbuted or not.

Lets check if our three numerical variables are normally distributed:

#weight  
  
ggplot(tiger\_data, aes(x = weight\_kg))+geom\_histogram(col = "white")  
shapiro.test(tiger\_data$weight\_kg)  
  
  
# parasite load  
  
ggplot(tiger\_data, aes(x = parasite\_load))+geom\_histogram(col = "white")  
shapiro.test(tiger\_data$parasite\_load)  
  
  
# age  
  
ggplot(tiger\_data, aes(x = age\_years))+geom\_histogram(col = "white", bins =8 )  
shapiro.test(tiger\_data$age\_years)

#### Exercise 3

Based on the shapiro wilk tests and histograms, are weight, parasite load, and age normally distributed or not?

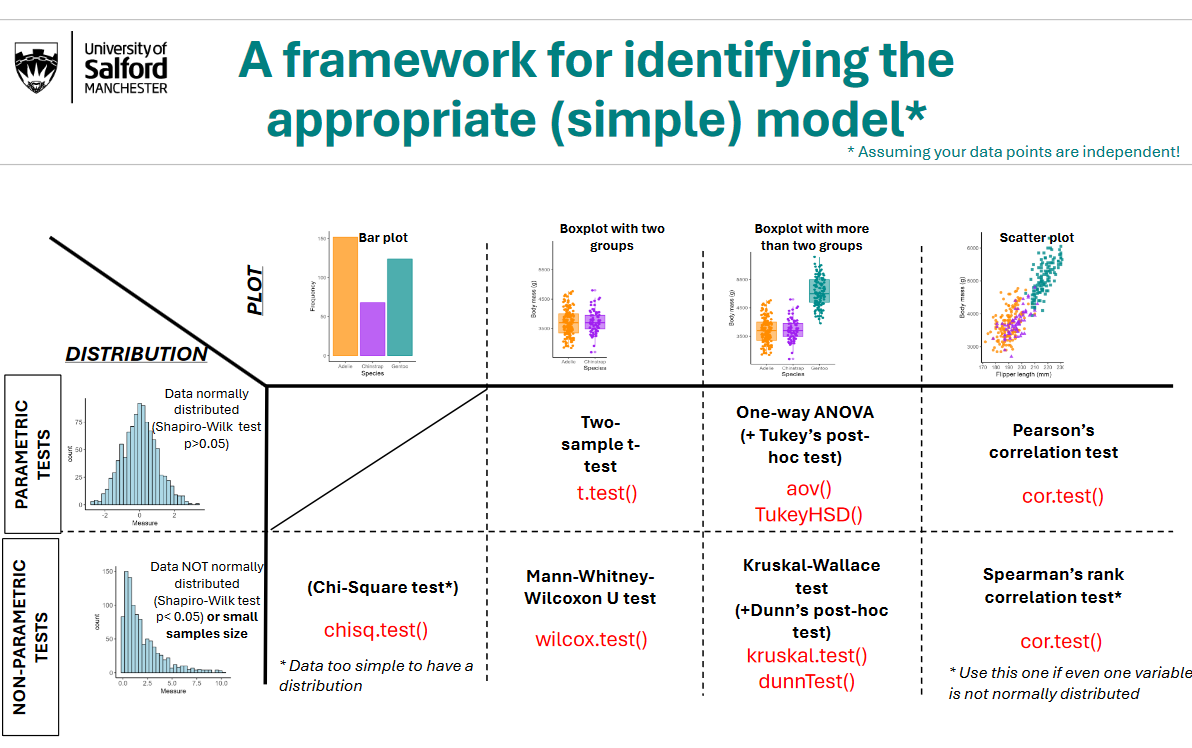
\*\* Write your answers below \*\*  
  
  
  
---

Lets next fit 5 univariate models to examine the effect of each variable on weight seperately.

#### Exercise 4

Use the shapiro wilk tests and the framework below to decide which test you need for each univariate association:

\*\* Write correct test below \*\*  
  
# Weight ~ Antibiotic treatment  
Correct test =   
  
# Weight ~ Diet treatment  
Correct test =   
  
# Weight ~ Age  
Correct test =   
  
# Weight ~ Parasite load  
Correct test =   
  
# Weight ~ Sex  
Correct test =   
  
---



#### Exercise 5

Using the code cheatsheet as a reference, conduct the five univariate statistical tests you identified in exercise 4. Write your conclusions based on the results of the univariate test, making sure to include your statistical evidence in brackets.

1. Weight ~ Antibiotic treatment

* \*\* Write your results below\*\*  
    
    
  ---

1. Weight ~ Diet treatment

\*\* Write your results below\*\*  
  
  
---

1. Weight ~ Age

\*\* Write your results below\*\*  
  
  
---

1. Weight ~ Parasite load

\*\* Write your results below\*\*  
  
  
---

1. Weight ~ Sex

\*\* Write your results below\*\*  
  
  
---

#### Exercise 6

Out of the five variables tested above, which had significant effects on weight, and which didn’t?

\*\* Write your answers below\*\*  
  
  
  
---

## Fitting a GLM

We have now looked at the relationship between weight and the five explanatory variables by plotting their relationships and fitting five univariate models.

However, this was a very long-winded approach! Lets now just fit one MULTIVARIATE general linear model with all variables in one model. We can then compare the results to our univariate models.

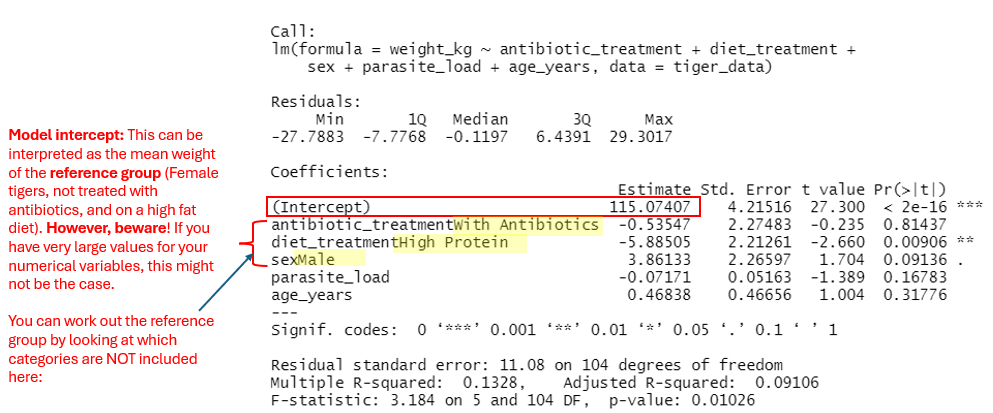
weight\_model <- lm(weight\_kg ~   
 antibiotic\_treatment +  
 diet\_treatment+  
 sex+  
 parasite\_load +   
 age\_years,   
 data = tiger\_data)  
  
summary(weight\_model)

## Interpreting a GLM output

There is a lot of break down in the output of a GLM. Below we will go through the output step by step.

### The intercept and reference groups

The first thing to look at is the first row of the Estimate column. This represents the INTERCEPT of your reference group (see below for explanation for what a reference group is). If you have no categorical variables in your model, this number will roughly represent the overall average of your response variable (in this case, the response variable is weight).



If you DO have numerical variables in your model (we do - age and parasite load), the intercept represents the estimated mean response when all predictor variables are set to zero.

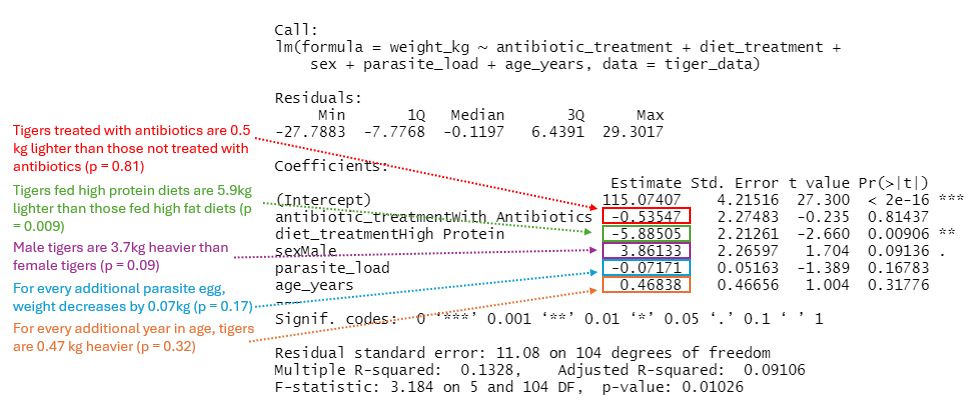
However, if your numerical variables are really large and not close to zero (e.g., blackbird weight in your assignment data), you might get a weird intercept value. This isn’t so important, it just means that number will be difficult to interpret.

**Reference group**: If you have categorical variables in your model, the model will automatically pick a reference group. In this case, the reference group of tigers is female, was not given atibiotics, and were fed a high fat diet.

Unless you reorder the levels manually, R automatically puts categorical levels in alphabetical order (e.g, female is before male when ordered alphabetically).

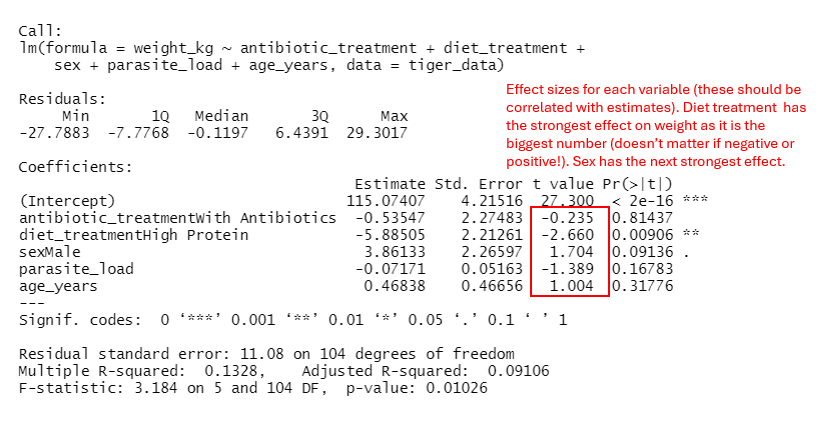
### Estimates

Estimates tell you the relationship between each explanatory variable and your response variable (in this case weight), in relation to the reference group. Below is a breakdown of what each estimate value means. Check out the p value to work out if the relationship is significant, and make sure you note whether the relationship is positive or negative.



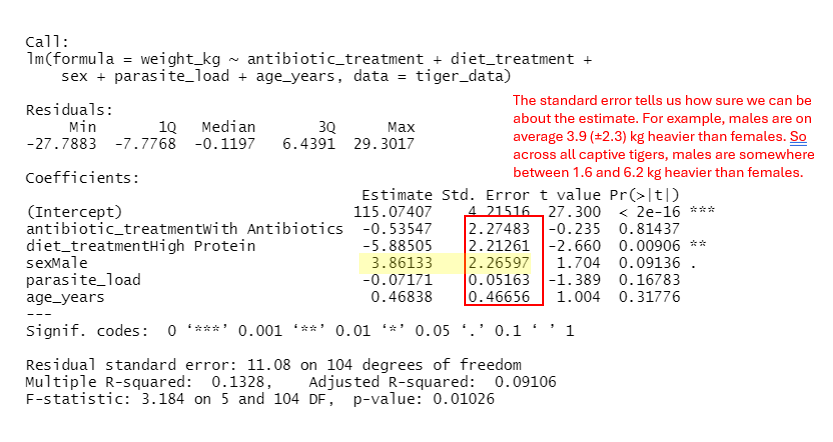
### Effect sizes - which variable has the strongest effect?

Which variables have the strongest and weakest effect on tiger weight? Look at the effect size!



### Standard errors (a measure of confidence in the estimate)

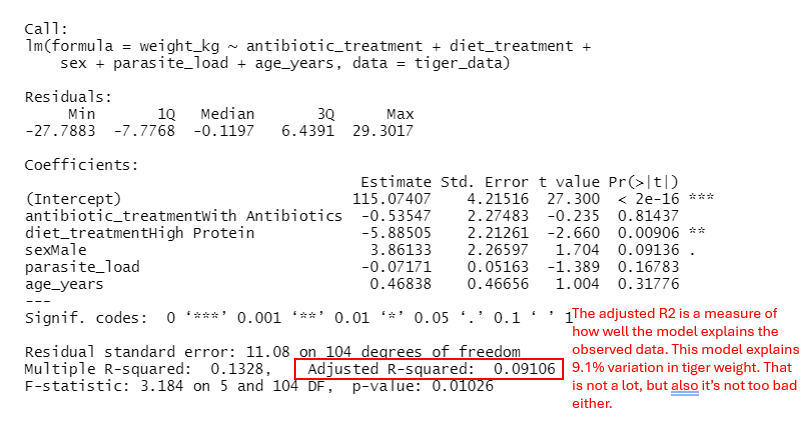
For which variable do we have the most confidence about the true estimate? Look at the standard error (or 95% confidence interval). Both standard errors and confidence intervals tell us something about the error around the estimate.



### Model fit

How good is our model at explaining tiger weight? The adjusted R2 tell your how well the model fits your data. In this case, the R2 is 0.09 - our explanatory variables in the model together account for 9% of the variation in tiger weight.

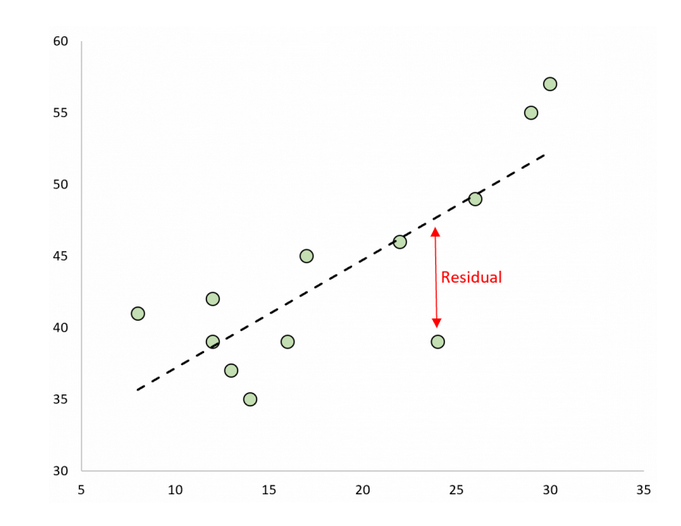
Given our R2 is quite low, there are probably other important variables out there that we havent measuresd and which are more important for explaining tiger weight than the variables we have included in our model.



### Model assumptions

The most important assumption of GLMs is that the RESIDUALS of the model are more or less normally distributed. Model residuals are the differences between the actual observed values and the predicted values in a statistical model.

In the figure below, the observed values are represented by the points, and the predicted values sit along the model regression line. Therefore, the residuals are the difference between the observed value and the predicted value:



I would normally recommend that you always use ggplot2 to plot your data. The one exception is if you want to quickly plot a histogram from a list of numbers. In this case, use the base function hist() - it is much simpler!

hist(residuals(weight\_model))

The residuals are normally distributed, so the main model assumption has been met. We can proceed to presenting the results of our model. If the residuals had been skewed, we would need to think about fitting a more complex model or transforming our data. We will not go into these problems here, but you may need to consider such problems when you are analysing your own data (e.g., for your thesis).

### Presenting GLM results as a table

Presenting model results can be tricky. There are two packages that help you tidy up the model output into a presentable format. Note that when presenting measures of confidence, you can either present the standard error OR 95% confidence interval. Both are acceptable.

# Use the tidy() function to extract the table  
broom::tidy(weight\_model)  
  
# You can save this table to your working directory as an excel file  
  
table<-tidy(weight\_model) # save table as an R object  
#write.csv(table, "weight\_model\_table.csv") # save file  
  
# OR zou can use the tab\_model() function in the R package sjPlot.   
# This function is great!  
  
sjPlot::tab\_model(weight\_model)  
  
  
tab\_model(weight\_model, show.se = T, show.stat = T) # Include standard error and effect size in output

### Presenting model results as a figure

It’s often good to present your model estimates as a figure because figures are always easier to digest than tables. You can use the function plot\_model() in the package sjPlot to plot your model estimates and confidence intervals.

sjPlot::plot\_model(weight\_model)  
  
# Add a line around zero. If the confidence intervals overlap zero, you know that variable is not a significant predictor of weight  
  
sjPlot::plot\_model(weight\_model)+  
 geom\_hline(yintercept = 0, linetype = "dashed") # add line

You can clearly see that diet treatment is by far the strongest predictor of tiger weight. Males are marginally heavier than females, but this effect is not significant (the confidence intervals do not overlap zero in the figure, AND the p value in the model is 0.09).

#### Exercise 7

How do the results from the GLM compare to those of your univariate tests? Are there any associations that differ in their significance?

\*\* Write your answers here \*\*  
  
  
  
  
---

### Writing conclusions based on GLMs

How would you write your conclusions? Remember, you are mainly interested in the effect of diet treatment so you should focus on this. Here is an example of how you could write out your results:

*“The weight of captive tigers was significantly associated with diet treatment, with tigers fed a high protein diet on average 5.9kg lighter than those fed a high fat diet (GLM, t = -2.7, p = 0.009). Antibiotic treatment had no impact on weight (GLM; t = -0.2, p = 0.81), nor did parasite load (GLM; t = -1.4, p = 0.17) and age (GLM; t = 1.0, p = 0.32). Males had a tendency to be heavier than females (GLM; t = 1.7, p = 0.09) but this association was not significant. The model explained over 9.1% of the variation in tiger weight (adjusted R2 = 0.091).”*

## Q2) Examining the predictors of penguin weight

Now it is your turn to explore the drivers of weight - but this time in penguins! You want to identify the strongest predictors of weight in three species of penguins in Antarctica.

Lets use an example dataset available from the R package palmerpenguins.

data(penguins) # load example dataset  
  
head(penguins) # check out dataset  
  
dim(penguins) # dataset consists of 344 penguins and 8 columns  
  
names(penguins) # column names  
  
str(penguins) # all looks ok apart from the varible year, which should be categorical, not numeric

#### Exercise 8

Plot the relationship between penguin weight (body\_mass\_g) and a) species, b) flipper\_length, c) bill length, and d) sex. Present the plots together using the ggarrange() function. Copy your composite plot below.

\*\* Copy plot below \*\*  
  
  
  
---

#### Exercise 9

Which variables look like they would be significantly associated with weight?

\*\* Copy your answer below \*\*  
  
  
  
---

#### Exercise 10

Fit a GLM that examines the effects of flipper length, bill length, species, and sex on penguin body mass.

\*\* Copy your code and output below \*\*  
  
  
---

#### Exercise 11

Are the residuals to the model normally distributed? Present the histogram below.

\*\* Copy the histogram below \*\*  
  
  
---

#### Exercise 12

Present a table of your model, including the standard errors and the effect size (effect size in GLMs = t value).

\*\* Copy the table below \*\*  
  
  
---

#### Exercise 13

Present a plot of your model estimates, using the function *plot\_model() .*

\*\* Copy the model below \*\*  
  
  
---

#### Exercise 14

Write out your results of your model, embedding statistical support from your model into your sentences. What are the strongest and weakers predictors of penguin weight? Which variables did not significantly predict weight? How much variation in penguin weight is explained by the model?

\*\* Copy your answer below \*\*  
  
  
---