Statistics for Psychologists - PSYC122

Margriet Groen and Rob Davies

2022-01-31

Contents

Chapter 1

Intro

This is a collection of tuition material written for Psychology undergraduates at Lancaster University. At the moment the content represents the "lab materials" for the PSYC122 module in first year. As was the case for PSYC121, they feature tuition of programming with R, building on the skills you developed last term.

1.1 Analysis labs and 'pre-lab work'

Some parts should be completed before you attend the lab session (watching lectures, reading chapters, pre-lab activities). All the links to the different materials and activities are also in the 'to-do list' for the relevant week on Moodle.

Chapter 2

Week 11: Correlation

Written by Margriet Groen (partly adapted from materials developed by the PsyTeachR team a the University of Glasgow)

Today we will take a look at correlation as a measure of association between two numerical variables. We will create scatterplots to visualise correlations, we will run a correlation analysis and we will practise interpreting and reporting the results.

2.1 Lectures

The lecture material for this week is presented in two parts:

1. **Theory** Watch this part before you complete the reading and the pre-lab activities.
2. **How to** Watch this part either after the 'Theory' part of after you've completed the pre-lab activities 1 to 3. Definitely watch it before you come to your lab session.

2.2 Reading

The reading that accompanies the lectures this week and next week is from **the free textbook by Miller and Haden**.

Chapter 10 gives you a brief overview of what correlation and regression are. **Chapter 11** introduces correlation in more detail. Both chapters are really short but provide a good basis to understanding correlational analysis. Please note, in Chapter 10 you might encounter some terminology that is unfamiliar

to you. It talks about ANOVA, which means Analysis of Variance and about GLM, which means General Linear Model. Having a quick look at Chapter 1 of Miller and Haden also helps with that.

2.3 Pre-lab activities

After having watched the lectures on correlation and read the textbook chapters you'll be in a good position to try these activities. Completing them before you attend your lab session will help you to consolidate your learning and help move through the lab activities more smoothly.

2.3.1 Pre-lab activity 1: Visualing correlations

Have a look at **this visualisation of correlations** by Kristoffer Magnusson.

After having read Miller and Haden Chapter 11, use this visualisation page to visually replicate the scatterplots in Figures 11.3 and 11.4 - use a sample of 100. After that, visually replicate the scatterplots in Figure 11.5.

Each time you change the correlation, pay attention to the shared variance (the overlap between the two variables) and see how this changes with the changing level of relationship between the two variables. The greater the shared variance, the stronger the relationship. Also, try setting the correlation to r = .5 and then moving a single dot to see how one data point, a potential outlier, can change the stated correlation value between two variables.

2.3.2 Pre-lab activity 2: Guess the correlation

Now that you are well versed in interpreting scatterplots (scattergrams) have a go at **this online app on guessing the correlation**.

This is a very basic app that allows you to see how good you are at recognising different correlation strengths from the scatterplots. We would recommend you click the "Track Performance" tab so you can keep an overview of your overall bias to underestimate or overestimate a correlation.

Is this all just a bit of fun? Well, yes, because stats is actually fun, and no, because it serves a purpose of helping you determine if the correlations you see in your own data are real, and to help you see if correlations in published research match with what you are being told. As you will have seen from the above examples, one data point can lead to a misleading relationship and even what might be considered a medium to strong relationship may actually have only limited relevance in the real world. One only needs to mention Anscombe's

Quartet to be reminded of the importance of visualising your data, which leads us to the final pre-lab activity for this week.

2.3.3 Pre-lab activity 3: Anscombe's quartet

Anscombe (1973) showed that four sets of bivariate data (X, Y) that have the exact same means, medians, and relationships can look very different when plotted. You can read more about this **here**.

All in this is a clear example of why you should visualise your data and not to rely on just the numbers.

2.3.4 Pre-lab activity 4: Getting ready for the lab class

2.3.4.1 Remind yourself of the basics of how to work with RStudio.

You might want to re-watch some of the videos John and Tom provided in PSYC121:

- Video on how to upload a zip file and import data (3.5 mins)
- Video on basic operations in RStudio (8 mins)
- Video on using scripts and using the console (3 mins)

2.3.4.2 Create a folder and a Project for Week 11.

Click here for the instructions from Week 6 of PSYC121 if you are unsure.

2.3.4.3 Get your files ready

Download the 122_week11_forStudents.zip file and upload it into the new folder in RStudio Server you created at the previous step. If you need them, here are the instructions from Week 2 of PSYC121.

2.4 Lab activities

In this lab, you'll gain understanding of and practice with:

- constructing and interpreting scatterplots
- running correlation analysis and interpret the results
- reporting the results in APA format
- constructing a correlation matrix in APA format

- when and why to apply correlation analysis to answers questions in psychological science

2.4.1 Lab activity 1: Interpreting correlation

2.4.1.1 Question 1

Below are scatterplots that show the relationship between 'how much you know about correlation and how attractive you appear to members of the opposite (&/or same) sex'. Choose the type of correlation (strength and direction) displayed in each graph using one of the following:

- Perfect positive correlation
- Perfect negative correlation
- Strong positive correlation
- Strong negative correlation
- Moderate positive correlation
- Moderate negative correlation
- Null correlation

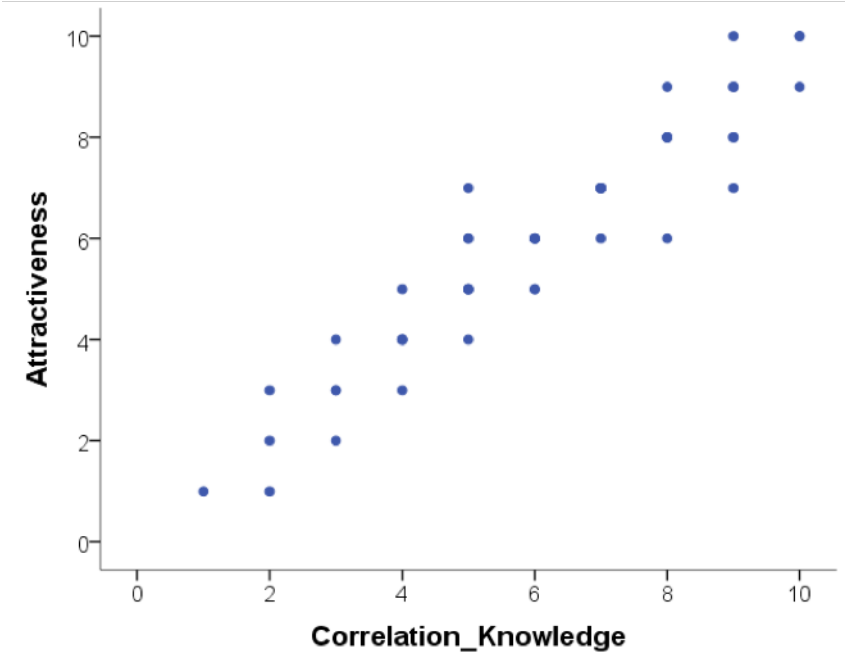


Figure A

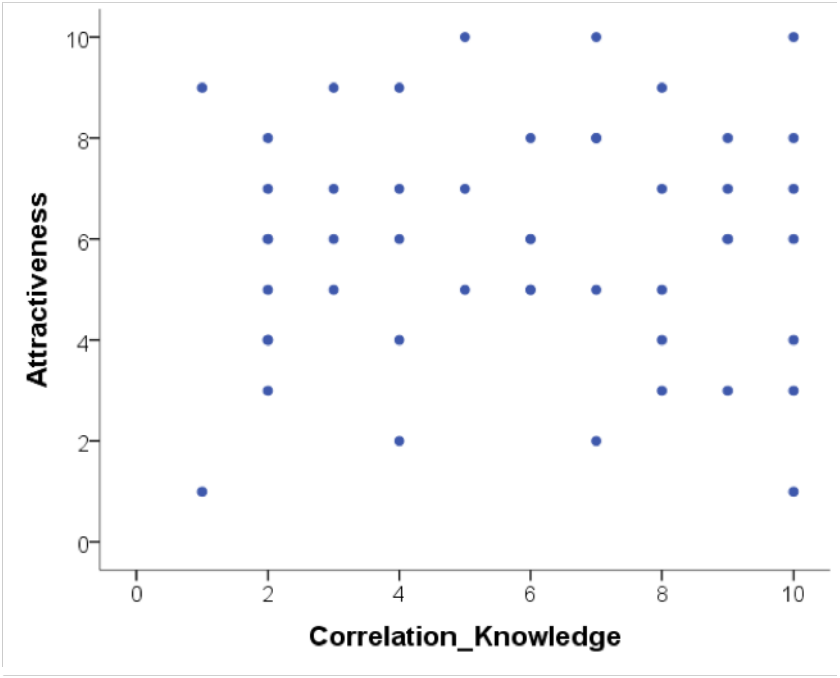


Figure B

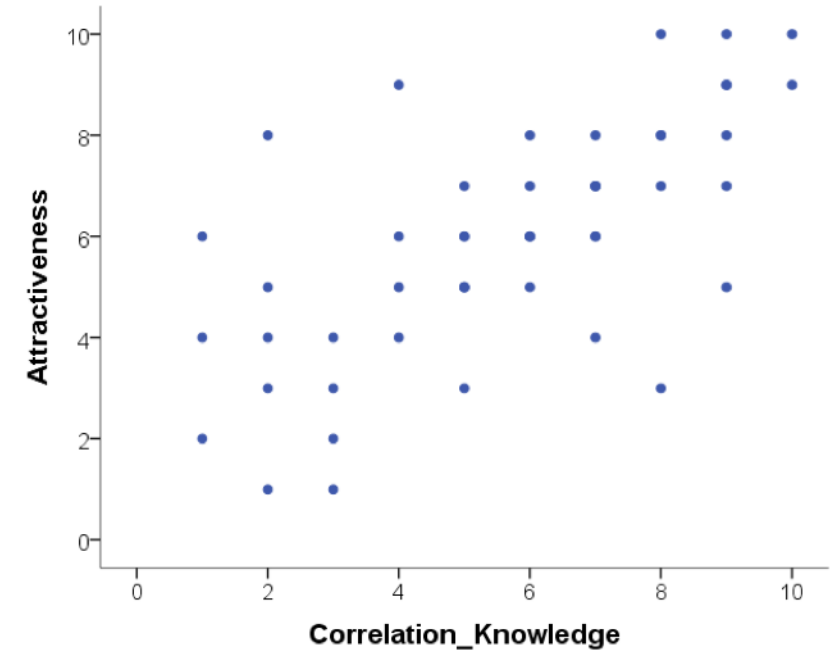


Figure C

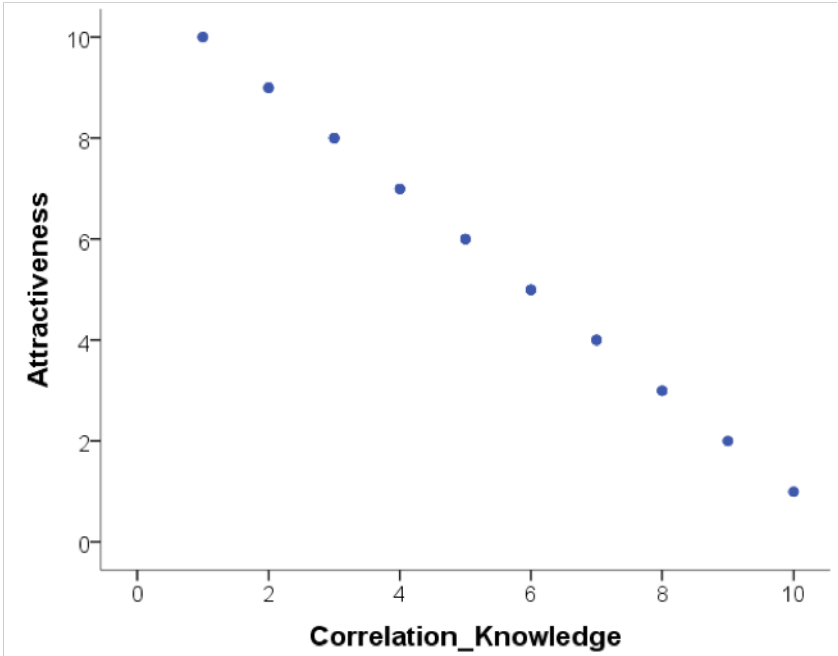


Figure D

2.4.1.2 Question 2

Suppose it was observed that there is a correlation of r = -.81 between a driver's age and the cost of car insurance. This correlation would mean that, in general, older people pay more for car insurance.

TRUE or FALSE? Explain why.

Note: explain your chosen answer based on the statistic given, not on why you think the correlation may or may note make 'logical' sense).

2.4.1.3 Question 3

Suppose that there is a correlation of r = .87 between the length of time a person is in prison and the amount of aggression the person displays on a psychological inventory administered at release. This means that spending a longer amount of time in prison *causes* people to become more aggressive.

TRUE or FALSE? Explain why.

2.4.1.4 Question 4

A significant correlation was found between having great hair and performance in correlation labs. The correlation coefficient was .7. How much variance in correlation lab performance can the 'greatness' of your hair explain?

- 51%
- 70%
- 49%
- 30%
- Who cares I've got great hair.

What was the reason for your answer?

What is this 'new coefficient' called?

2.4.2 Lab activity 2: Visualising, calculating and reporting correlations

Going back to the data discussed in Chapter 11 of Miller & Haden, you'll remember it contains data from 25 8-year-old children on:

- a standardised test of reading ability (Abil)
- intelligence (IQ)
- the number of minutes per week spent reading in the home (Home)
- and the number of minutes per week spent watching TV (TV)

In the video on 'How to conduct correlation analysis using R' we looked at the correlation between reading ability and intelligence. Now, let's look at the correlation between number of minutes per week spent reading in the home and watching TV.

The folder you were asked to download under 'Pre-lab activity 4: Getting ready for the lab class' contains the datafile ("MillerHadenData.csv") as well as the R-script from the 'How to …' video (122_wk11_howtoExample.R) that you can use here and adapt.

1. Load the 'broom' and the 'tidyverse' libraries by running the first two lines of code.
2. Read in the data. You should now see an object with 25 observations and 5 variables in the 'Environment'. Click on it to view it.
3. Construct a scatterplot of the relationship between 'Home' and 'TV'.
4. What can you tell from the scatterplot about the direction of the relationship?
5. Conduct the correlation analysis.

6. What is the correlation coefficient (Pearson's r)?
7. What is the p value?
8. Is the correlation significant at the $p < .05$ level?
9. What are the degrees of freedom you need to report?
10. How much variance in 'time spent reading' can be accounted for by 'time spent watching TV'? (Hint: you can use the Console in RStudio as a calculator.)
11. Write a few sentences in which you report this result, following APA guidelines.

2.4.3 Lab activity 3: More correlations

Researchers were interested in the relationship between hazardous alcohol use and impulsivity (making unplanned, rapid decisions without thinking or 'acting on a whim'). To investigate the relationship, 20 participants completed both the alcohol use disorder identification test (AUDIT; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993) and the Barratt's Impulsiveness Scale (BIS-11) (Patton, Stanford, & Barratt, 1995). The datafile ("alcoholUse_Impulsivity.csv") is in the folder you were asked to download under 'Pre-lab activity 4: Getting ready for the lab class'. Again, the R-script from the 'How to …' video (122_wk11_howtoExample.R) is useful here.

1. Load the 'broom' and the 'tidyverse' libraries by running the first two lines of code.
2. Read in the data. You should now see an object containing the data in the 'Environment'. How many variables does it have?
3. Construct a scatterplot of the relationship between 'Hazardous Alcohol Use' and 'Impulsivity'.
4. What can you tell from the scatterplot about the direction of the relationship?
5. Conduct the correlation analysis.
6. What is the correlation coefficient (Pearson's r)?
7. What is the p value?
8. Is the correlation significant at the $p < .05$ level?
9. What are the degrees of freedom you need to report?
10. How much variance in 'impulsivity' can be accounted for by 'hazardous alcohol use'? (Hint: you can use the Console in RStudio as a calculator.)

11. Construct a correlation matrix to display the correlation coefficient in a table.
12. Give three logically possible directions of causality, indicating for each direction whether it is a plausible explanation in light of the variables involved (and why). No, this is not a trick question -—I know that correlation does not infer causation, but think critically! New studies/ideas are constructed by thinking what the previous study doesn't tell us about what could be happening with the variables of interest.

Job completed — Well done!

2.5 Answers

When you have completed all of the lab content, you may want to check your answers with our completed version of the script for this week. **Remember**, looking at this script (studying/revising it) does not replace the process of working through the lab activities, trying them out for yourself, getting stuck, asking questions, finding solutions, adding your own comments, etc. **Actively engaging** with the material is the way to learn these analysis skills, not by looking at someone else's completed code…

2.5.1 Lab activity 1: Interpreting correlation

1. Scatterplots

- a. strong positive correlation
- b. null correlation
- c. moderate positive correlation
- d. perfect negative correlation

2. FALSE Explanation: The correlation coefficient is negative and therefore infers a negative correlation. As such, older people pay less for car insurance: as age increases, car insurance costs decrease.
3. FALSE Explanation: This is a bit of trick question as it has the sneaky 'cause' word in. The correlation coefficient is a positive number, suggesting a positive relationship between length of time in prison and aggression. However, causation cannot be inferred from correlation and therefore we cannot know whether time spent in prison CAUSES aggression, and rather we suggest a relationship between the two that as length of time in prison increases, aggression increases.
4. c 49% The 'coefficient of determination' or 'R-squared' tells us the proportion or variance in one variable that can be predicted if we know the

other variable. We can determine this by squaring the r. Therefore, $.7^2 = .49$, $R^2 = .49$.

2.5.2 Lab activity 2: Constructing scatterplots and calculating correlations

You can download the R-script that contains the code to complete lab activities 2 and 3 here: 122_wk11_labActivities2_3.R.

1. *See R script*
2. *See R script*
3. *See R script*
4. What can you tell from the scatterplot about the direction of the relationship? **There is a negative association between 'Home' and 'TV'. This means that the longer a child spends watching TV, the shorter they will read at home.**
5. Conduct the correlation analysis. *See R script*
6. What is the correlation coefficient (Pearson's r)? $\boldsymbol{r = -.65}$
7. What is the p value? $\boldsymbol{p < .001}$
8. Is the correlation significant at the p < .05 level? **Yes, because the p-value is smaller than .005**
9. What are the degrees of freedom you need to report? **23**
10. How much variance in 'time spent reading' can be accounted for by 'time spent watching TV'? **42%**
11. Write a few sentences in which you report this result, following APA guidelines. **Something along the lines of: A Pearson's correlation coefficient was used to assess the relationship between time spent reading at home and time spent watching TV at home. There was a significant negative correlation, $\boldsymbol{r(23) = -.65}$, $\boldsymbol{p < .001}$. As time spent watching TV at home increased, time spent reading at home decreased.**

2.5.3 Lab activity 3: More correlations

1. *See R script*
2. How many variables does it have? **3**
3. *See R script*

4. What can you tell from the scatterplot about the direction of the relationship? **There is a positive association between 'hazardous alcohol use' and 'impulsivity'. This means that as a participant's score on 'hazardous alcohol use' goes up, their score on 'impulsivity' also goes up.**
5. *See R script*
6. What is the correlation coefficient (Pearson's r)? $\boldsymbol{r = .54}$
7. What is the p value? $\boldsymbol{p = .014}$
8. Is the correlation significant at the $p < .05$ level? **Yes**
9. What are the degrees of freedom you need to report? **18**
10. How much variance in 'impulsivity' can be accounted for by 'hazardous alcohol use'? (Hint: you can use the Console in RStudio as a calculator.) **29%**
11. Construct a correlation matrix to display the correlation coefficient in a table.

Table 1. ***A correlation matrix showing the relationship between hazardous alcohol use and impulsivity.***

	Hazardous alcohol use	Impulsivity
Hazardous alcohol use	-	
Impulsivity	.54*	-

***p < .05**

12. Give three logically possible directions of causality, indicating for each direction whether it is a plausible explanation in light of the variables involved (and why). No, this is not a trick question -—I know that correlation does not infer causation, but think critically! New studies/ideas are constructed by thinking what the previous study doesn't tell us about what could be happening with the variables of interest.

Just really looking for reasoning here.

Examples:

- **Being more impulsive may make people consume more alcohol.**
- **Consuming more alcohol may make people more impulsive.**
- **An outgoing personality might influence both your level of impulsivity and you are more likely to be socialising in the pub and consuming alcohol. So the same 'third factor' may influence both our variables of interest.**

Chapter 3

Week 12: Correlation 2

Written by Margriet Groen (partly adapted from materials developed by the PsyTeachR team a the University of Glasgow)

Today we will continue a look at correlation as a measure of association between two numerical variables. We will review assumptions associated with correlation, discuss some issues important to be aware of when interpreting correlation results and finally, we'll talk about intercorrelation.

3.1 Lectures

The lecture material for this week is presented in two parts:

1. **Correlation – Assumption, issues and intercorrelation – Theory**
2. **Correlation – Assumption, issues and intercorrelation – How to**

3.2 Reading

The reading that accompanies the lectures this week is (the same as last's week) from **the free textbook by Miller and Haden**.

Chapter 10 gives you a brief overview of what correlation and regression are. **Chapter 11** introduces correlation in more detail. Both chapters are really short but provide a good basis to understanding correlational analysis. Please note, in Chapter 10 you might encounter some terminology that is unfamiliar to you. It talks about ANOVA, which means Analysis of Variance and about GLM, which means General Linear Model. Having a quick look at Chapter 1 of Miller and Haden also helps with that.

3.3 Pre-lab activities

After having watched the lectures on correlation and read the textbook chapters you'll be in a good position to try these activities. Completing them before you attend your lab session will help you to consolidate your learning and help move through the lab activities more smoothly.

3.3.1 Pre-lab activity 1: Online interactive tutorial to practise your data-wrangling skills

Data comes in lots of different formats. One of the most common formats is that of a two-dimensional table (the two dimensions being rows and columns). Usually, each row stands for a separate observation (e.g. a participant), and each column stands for a different variable (e.g. a response, category, or group). A key benefit of tabular data is that it allows you to store different types of data-numerical measurements, alphanumeric labels, categorical descriptors-all in one place.

It may surprise you to learn that scientists actually spend far more of time cleaning and preparing their data than they spend actually analysing it. This means completing tasks such as cleaning up bad values, changing the structure of tables, merging information stored in separate tables, reducing the data down to a subset of observations, and producing data summaries. Some have estimated that up to 80% of time spent on data analysis involves such data preparation tasks (Dasu & Johnson, 2003)!

Many people seem to operate under the assumption that the only option for data cleaning is the painstaking and time-consuming cutting and pasting of data within a spreadsheet program like Excel. We have witnessed students and colleagues waste days, weeks, and even months manually transforming their data in Excel, cutting, copying, and pasting data. Fixing up your data by hand is not only a terrible use of your time, but it is error-prone and not reproducible. Additionally, in this age where we can easily collect massive datasets online, you will not be able to organise, clean, and prepare these by hand.

In short, you will not thrive as a psychologist if you do not learn some key data wrangling skills. Although every dataset presents unique challenges, there are some systematic principles you should follow that will make your analyses easier, less error-prone, more efficient, and more reproducible.

In the online tutorial, you will see how data science skills will allow you to efficiently get answers to nearly any question you might want to ask about your data. By learning how to properly make your computer do the hard and boring work for you, you can focus on the bigger issues.

You'll be practising the `select()`, `filter()`, `mutate()`, `arrange()`, `group_by()` and `summarise()` functions from the dplyr package.

You've used these functions before, but if you'd like to quickly remind yourself what they do, watch the video (~10 min) on **Data wrangling: dplyr and pipes**. As the title suggests, I also explain in the video what a 'pipe' (this thing: `%>%`) is and you'll be practising with that as well.

If you're ready to begin, go to the tutorial linked to below. There is no need to install or download anything. Each tutorial has everything you need to write and run R code, right in the tutorial.

- **Working with Tibbles** Practise how to extract values form a table, subset tables, calculate summary statistics, and derive new variables.

3.3.2 Pre-lab activity 2: Getting ready for the lab class

3.3.2.1 Get your files ready

Download the 122_week12_forStudents.zip file and upload it into the new folder in RStudio Server you created (see last week's Pre-lab activity 4 for instructions on how to do that.

3.4 Lab activities

In this lab, you'll gain understanding of and practice with:

- constructing and interpreting histograms and qq-plots
- constructing and interpreting a matrix of scatterplots
- running intercorrelation analysis and interpret the results
- correct for multiple comparisons when running intercorrelation analysis
- constructing a correlation matrix in APA format
- when and why to apply correlation analysis to answers questions in psychological science

3.4.1 Lab activity 1: Assumptions of Correlation Analysis

3.4.1.1 Question 1

Correlation would be an appropriate form on analysis for researchers interested in the relationship between:

a. Dog (breed) and height (cm) of owner
b. Speed of swimming (mph) and area of tank (cm)
c. Number of cows sitting and rain fall (mm)
d. Total llama saliva (ml) expelled and gender of visitors
e. b and d

f. b and c

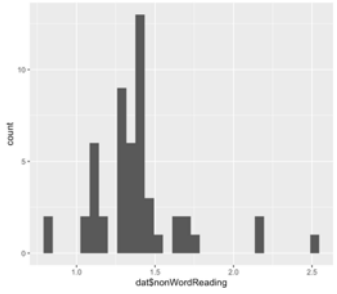
3.4.1.2 Question 2

When would you use Spearman's rho analysis instead of Pearson's r?

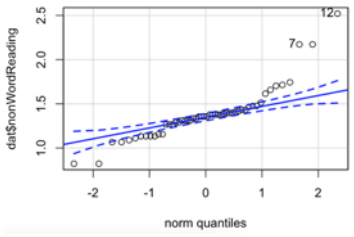
a. When there are clear outliers in the data
b. When the data is not normally distributed
c. When the relationship between X and Y is curvilinear
d. a and b

3.4.1.3 Question 3

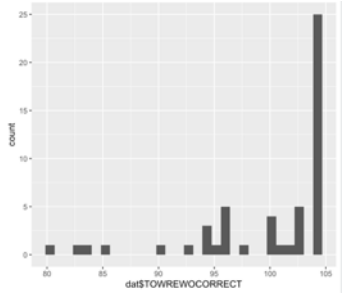
Using the histograms and qq-plots below, which of these variables satisfies the normality assumption? Explain your answers.



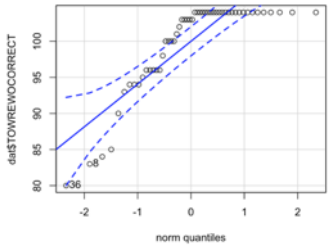
Histogram non-words



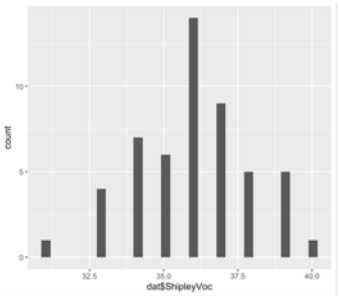
QQ-plot non-words



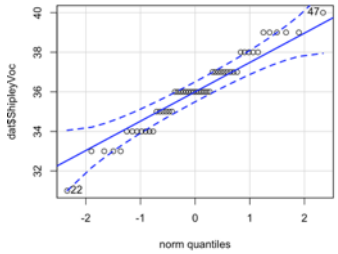
Histogram words



QQ-plot words



Histogram vocabulary



QQ-plot non-words

3.4.1.4 Question 4

Why should correlation analysis not be conducted on variables with a curvilinear relationship?

3.4.2 Lab activity 2: Attitudes towards vaping

Great work so far! Now we really want to see what you can do yourself. In this activity we'll use real data on implicit and explicit attitudes towards vaping. You'll need the data file `VapingData.csv` and the R-script `122_wk12_labAct2_template.R` that you downloaded when completing Pre-lab activity 2.

3.4.2.1 Background

Explicit attitudes were measured via a questionnaire where higher scores indicated a positive attitude towards vaping (`VapingQuestionaireScore`). **Implicit attitudes** were measured through an Implicit Association Test (IAT) using images of Vaping and Kitchen utensils and associating them with positive and negative words.

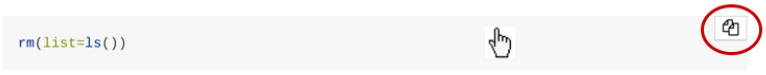
The IAT works on the principal that associations that go together (that are congruent, e.g. warm and sun) should be quicker to respond to than associations that do not go together (that are incongruent, e.g. warm and ice). You can read up more on the procedure on **the Noba Project** which has a good description of the procedure under the section "Subtle/Nonsconscious Research Methods".

For this exercise, you need to know that "Block 3" in the experiment tested reaction times and accuracy towards congruent associations, pairing positive words with Kitchen utensils and negative words with Vaping. "Block 5" in the experiment tested reaction times and accuracy towards incongruent associations, pairing positive words with Vaping and negative words with Kitchen Utensils. As such, if reaction times were longer in Block 5 than in Block 3 then people are considered to hold the view that Vaping is negative (i.e. congruent associations are quicker than incongruent associations). However, if reaction times were quicker in Block 5 than in Block 3 then people are considered to hold the view that Vaping is positive (i.e. incongruent associations were quicker than congruent associations). The difference between reaction times in Block5 and Block3 is called the participants' IAT score.

3.4.2.2 Step 0: Clean your environment

Before you do anything else, when starting a new analysis, it is a good idea to empty the R environment. This prevents objects and variables from previous analyses interfering with the current one. You can do this by clicking on the broom icon at the top of the environment window, or you can use the code below.

> **TASK**: Use the code snippet below to clear the environment. **TIP**: If you hover your mouse over the box that includes the code snippet, a 'copy to clipboard' icon will appear in the top right corner of the box. Click that to copy the code. Now you can easily paste it into your script.




```
rm(list=ls())
```

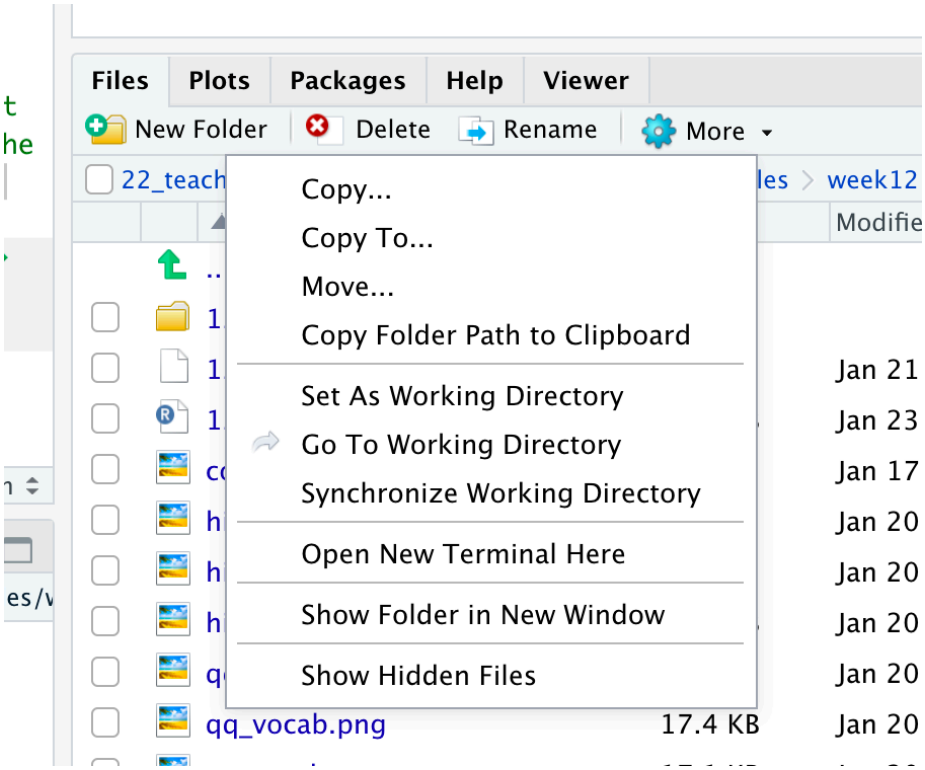
3.4.2.3 Step 1: Set your working directory

Now, make sure your working directory is set to the folder in which you have stored the data file (`VapingData.csv`).

> **TASK**: Use the code snippet below to check what you working directory is currently set to. This is the folder that R will use to look for files. Is the file path that is written to the Console after you run the code snippet the one that contains the data file? You can check by nativating to the path you can see in the Console in the 'Files' pane on the right. Does it contain the data file ('VapingData.csv')?

```
getwd()
```

If your working directory is not set to the folder that contains the data file, navigate to folder that contains the data file in the 'Files' pane, click 'More' and then on 'Set as working directory'.



3.4.2.4 Step 2: Load packages

Before we can get started we need to tell R which libraries to use. For this analysis we'll need `broom`, `car`, `Hmisc`, `lsr` and `tidyverse`.

> **TASK**: Load the relevant libraries. **HINT**: Use the `library()` function.

3.4.2.5 Step 3: Read in the data

The data file we'll be working with is `VapingData.csv`

> **TASK**: Read in the data file (using the `read_csv()` function) and store it in an object called `dat`. Have a look at the layout of the data and familiarise yourself with it. You have 8 columns. Reaction times and Accuracy scores for Blocks 3 and 5 as well as the Explicit Vaping Questionnaire scores, Sex and Age, for each participant.

QUESTION 1: For how many participants do we have data?

3.4.2.6 Step 4: Data wrangling

The data are not in a shape yet that we can actually use for our analysis. We'll have to do some 'data wrangling to knock them into shape. We need to take the following into account:

1. Accuracy was calculated as proportion and as such can't go above 1.Participants entered their own data so some might have made a mistake. Remove participants who had an accuracy greater than 1 in either Block 3 or Block 5 as we are unclear on the accuracy of these values.
2. We also only want participants who were paying attention so best remove anybody whose average accuracy score across Blocks 3 and 5 was less than 80%. Note - this value is arbitrary and if you wanted, in your own experiment, you could use a more relaxed or strict cut-off based on other studies or guidance. Note that these decisions should be set out at the start of your research as part of your pre-registration or as part of your Registered Report. Finally, in this instance, remember, the values are in proportions not percentages (so 80% will be .8).
3. Now that we have removed data points that were the result of data entry mistakes or came participants who didn't pay attention during the task, we create an IAT score for participants by subtracting Block 3 reaction times (RT) from Block 5 reaction times (IAT_BLOCK5_RT - IAT_BLOCK3_RT).

TASK: Look closely at each line of code below and check you understand what it does. Copy the code to your script and for each line add a comment to describe what it does.

```
dat <- dat %>%
  filter(IAT_BLOCK3_Acc < 1) %>%
  filter(IAT_BLOCK5_Acc < 1) %>%
  mutate(IAT_ACC = (IAT_BLOCK3_Acc + IAT_BLOCK5_Acc)/2) %>%
  filter(IAT_ACC > .8) %>%
  mutate(IAT_RT = IAT_BLOCK5_RT - IAT_BLOCK3_RT)
```

QUESTION 2: For how many participants do we have data now that we have cleaned them up?

QUESTION 3: Use the information in the background description to understand how the scores relate to attitudes. What does a positive IAT_RT score reflect? What does a negative IAT_RT score reflect? What does a higher score on the 'VapingQuestionnaireScore' mean?

3.4.2.7 Step 5: Calculating descriptive statistics

Now that we have the variables that we need and the data cleaned up, we will create a descriptives summary of the number of people, and the means for the IAT and Vaping Questionnaire Score.

TASK: Look closely at each line of code below and check you understand what it does. Copy the code to your script and for each line add a comment to describe what it does.

```
descriptives <- dat %>%
  summarise(n = n(),
            mean_IAT_ACC = mean(IAT_ACC),
            mean_IAT_RT = mean(IAT_RT),
            mean_VPQ = mean(VapingQuestionnaireScore, na.rm = TRUE))
```

QUESTION 4: Why might these averages be useful? Why are averages not always useful in correlations?

3.4.2.8 Step 6: Check the assumptions

Variable types

QUESTION 5: What are the variable types for the implicit (IAT_RT) and the explicit (VapingQuestionnaireScore) attitude variables?

Missing data

> **TASK**: Remove participants with missing data. **HINT**: Use the `filter()` function and the `is.na()` function

QUESTION 6: How many people had missing data?

Normality

> **TASK**: Create histograms and qq-plots for the `IAT_RT` and `VapingQuestionnaireScore` variables. **HINT**: Use the `ggplot()` function with `geom_histogram()` and use the `qqPlot()` function (note the capital P)

QUESTION 7: What do you conclude from the histograms and the qq-plots? Are the VapingQuestionnaireScore and the IAT_RT normally distributed?

Linearity and homoscedasticity

> **TASK**: Plot the relationship between IAT_RT and VapingQuestionnaireScore using a scatterplot and a line of best fit. **HINT**: For this you'll need the `ggplot()` function together with `geom_point()` and `geom_smooth()`. Make sure to give your axes some sensible labels.

QUESTION 8: What do you conclude from the scatterplot in terms of the homoscedasticity of the data and the linearity, direction and strength of the relationship? What does the scatterplot tell you about possible issues (outliers, range restrictions)?

3.4.2.9 Step 7: Conduct a correlation analysis

QUESTION 9: Do you need to calculate Pearson's r or Spearman's rho?

> **TASK**: Conduct a correlation analysis. **HINT**: Use the `cor.test()` function. You may want to use the `pull()` and the `round()` functions to get the numbers out.

QUESTION 10: Can you write up the results including the r, df, p-value and an interpretation?

3.4.2.10 Step 8: Intercorrelations

Finally, let's check whether either implicit or explicit attitude is associated with age.

First, let's create a new data frame that only includes the relevant variables. Look closely at each line of code below and check you understand what it does. Don't forget to copy the code below to your script and run it.

```
dat_matrix <- dat %>%
  select(Age, IAT_RT, VapingQuestionnaireScore) %>%
  as.data.frame(dat_matrix) # Make sure tell R that dat is a data frame
```

> **TASK**: Now, create a matrix of scatterplots. **HINT**: Use the `pairs()` function.

QUESTION 11: What do you conclude from the scatterplots?

> **TASK**: Finally, conduct intercorrelation (multiple correlations). **HINT**: Use the `correlate()` function. Do you need Pearson's *r* or Spearman's *rho*? Have you adjusted for multiple comparisons?

QUESTION 12: What do you conclude from the results of the correlation analysis?

3.5 Answers

When you have completed all of the lab content, you may want to check your answers with our completed version of the script for this week. **Remember**, looking at this script (studying/revising it) does not replace the process of working through the lab activities, trying them out for yourself, getting stuck, asking questions, finding solutions, adding your own comments, etc. **Actively engaging** with the material is the way to learn these analysis skills, not by looking at someone else's completed code…

3.5.1 Lab activity 1

1. Correlation would be an appropriate form on analysis for researchers interested in the relationship between

- a. Dog (breed) and height (cm) of owner
- b. Speed of swimming (mph) and area of tank (cm)
- c. Number of cows sitting and rain fall (mm)
- d. Total llama saliva (ml) expelled and gender of visitors
- e. b and d
- **f. b and c All variables in b anc c are continuous. Dog breed and gender are categorical.**

2. When would you use Spearman's rho analysis instead of Pearson's r?

- a. When there are clear outliers in the data
- **b. When the data is not normally distributed**
- c. When the relationship between X and Y is curvilinear
- d. a and b

3. Using the histograms and qq-plots below, which of these variables satisfies the normality assumption? Explain your answers. **Vocabulary. Only for vocabulary does the histogram resemble a bell curve and do the data-points in the qq-plot fall within the dashed blue lines.**

4. Why should correlation analysis not be conducted on variables with a curvilinear relationship? **May be subject to a type 2 error there actually is a relationship between variables yet we reject the null hypothesis. As the relationship is not linear, correlation analysis will not identify this.**

3.5.2 Lab activity 2

You can download the R-script that contains the code to complete lab activity 2 here: 122_wk12_labAct2.R.

1. For how many participants do we have data? **There are 166 observations, so we have data for 166 participants. You can see this in the Environment window in the top right. This does not tell us whether any of these participants have any missing data.**

2. For how many participants do we have data now that we have cleaned them up? **104 participants**

3. Use the information in the background description to understand how the scores relate to attitudes. What does a positive IAT_RT score reflect? **People with a positive IAT_RT are considered to hold the implicit view that vaping is negative (i.e. congruent associations are quicker than incongruent associations)** What does a negative IAT_RT score reflect? **People with a negative IAT_RT are considered to hold the implicit view that vaping is positive (i.e. incongruent associations were quicker than congruent associations).** What does a higher score on the 'VapingQuestionnaireScore' mean? **Higher scores indicated a positive explicit attitude towards vaping.**

4. Why might these averages be useful? Why are averages not always useful in correlations? **It is always worth thinking about which averages are informative and which are not. Knowing the average explicit attitude towards vaping could well be informative. In contrast, if you are using an ordinal scale and people use the whole of the scale then the average may just tell you the middle of the scale you are using - which you already know and really isn't that informative. So it is always worth thinking about what your descriptives are calculating.**

5. What are the variable types for the implicit (IAT_RT) and the explicit

(VapingQuestionnaireScore) attitude variables? **Both can be considered continuous variables and at least at interval level.**

6. How many people had missing data? **8. Before we removed participants with missing data, we had 104 observations, now we have 96. So there must have been 8 participants without a score on one or the other variable.**

7. What do you conclude from the histograms and the qq-plots? Are the VapingQuestionnaireScore and the IAT_RT normally distributed? **Yes. Both histograms resemble a normal distribution (bell curve) and the open circles in the qq-plots fall within the blue stripy lines.**

8. What do you conclude from the scatterplot in terms of the homoscedasticity of the data and the linearity, direction and strength of the relationship? What does the scatterplot tell you about possible issues (outliers, range restrictions)? **The data look like a cloud without a clear direction. This suggests the relationship might might be weak. In terms of linearity, the scatterplot doesn't suggest any curvilinear relationships. Variance seems quite constant, but there do seem to be few people with negative IAT_RT (Implicit attitude) scores, suggesting few people held the view that vaping is positive.**

9. Do you need to calculate Pearson's r or Spearman's rho? **Pearson's r because the data do meet the assumptions.**

10. Can you write up the results including the r, df, p-value and an interpretation? **Testing the hypothesis of a relationship between implicit and explicit attitudes towards vaping, a Pearson correlation found no significant relationship between IAT reaction times (implicit attitude) and answers on a Vaping Questionnaire (explicit attitude), $r(94) = -.02$, $p = .822$. Overall this suggests that there is no direct relationship between implicit and explicit attitudes with regard to vaping and as such our hypothesis was not supported; we cannot reject the null hypothesis.**

11. What do you conclude from the scatterplots? **The scatterplots with age suggest that age is highly skewed with only a few participants older than 25. For now, let's say we'll therefore calculate Spearman's rho, rather than Pearson's *r*. That is ok for now, but if you were analysing these data for a research project, you'd want to have a closer look at the age variable (think histogram, qq-plot, and think about either collecting more data from older participants or transforming the variable (more about that next year).**

12. What do you conclude from the results of the correlation analysis? **No significant correlation with age was found.**

Chapter 4

Week 13: The linear model

Written by Margriet Groen (partly adapted from materials developed by the PsyTeachR team a the University of Glasgow)

This week we will focus on the linear model and simple linear regression.

4.1 Lectures

The lecture material for this week is presented in two parts:

1. **The linear model (~25 min)**
2. **How to build a linear model in R (~30 min)** You can find the example script in this week's zip-folder (see under Pre-lab activity 3).

4.2 Reading

The reading that accompanies the lectures this week is from **the free textbook by Miller and Haden**.

Chapter 12 provides an accessible overview of simple regression.

4.3 Pre-lab activities

After having watched the lectures and read the textbook chapter you'll be in a good position to try these activities. Completing them before you attend your lab session will help you to consolidate your learning and help move through the lab activities more smoothly.

4.3.1 Pre-lab activity 1: Visualising the regression line

Have a look at **this visualisation of the regression line** by Ryan Safner.

In this shiny app, you see a randomly-generated set of data points (within specific parameters, to keep the graph scaled properly). You can choose a slope and intercept for the regression line by using the sliders. The graph also displays the residuals as dashed red lines. Moving the slope or the intercept too much causes the generated line to create much larger residuals. The shiny app also calculates the sum of squared errors (SSE) and the standard error of the regression (SER), which calculates the average size of the error (the red numbers). These numbers reflect how well the regression line fits the data, but you don't need to worry about those for now.

In the app he uses the equation Y = aX + b in which b is the intercept and a is the slope.

This is slightly different from the equation you saw during the lecture. There we talked about Y = b0 + b1*X + e. Same equation, just different letters. So b0 in the lecture is equivalent to b in the app and b1 in the lecture is equivalent to a in the app.

Pre-lab activity questions:

1. Change the slider for the intercept. How does it change the regression line?
2. Change the slider for the slope. How does it change the regression line?
3. What happens to the residuals (the red dashed lines) when you change the slope and the intercept of the regression line?

4.3.2 Pre-lab activity 2: Data visualisation - practice with `ggplot2()`

In this week's online tutorials, you will practise visualing data.

If you're ready to begin, go to the tutorial linked to below. There is no need to install or download anything. Each tutorial has everything you need to write and run R code, right in the tutorial.

- **Visualisation basics** Practise the basics of how to create a graph, how to add variables and how to make different types of graphs.
- **Scatterplots** This tutorial revisits scatterplots. Along the way, you will learn to build multi-layer plots and to use new coordinate systems.

4.3.3 Pre-lab activity 3: Getting ready for the lab class

4.3.3.1 Get your files ready

Download the 122_week13_forStudents.zip file and upload it into the new folder in RStudio Server you created (see week 12 Pre-lab activity 4 for instructions on how to do that).

4.4 Lab activities

In this lab, you'll gain understanding of and practice with:

- conducting simple regression in R
- interpreting simple regression in R
- reporting the results in APA format
- when and why to apply simple regression to answer questions in psychological science

4.4.1 Lab activity 1: The regression line

4.4.1.1 Question 1

What is the regression equation as discussed during the lecture and what does each letter represent?

4.4.1.2 Question 2

What are residuals?

4.4.1.3 Question 3

Discuss the answers to the pre-lab activity questions. What did you find?

a) Change the slider for the intercept. How does it change the regression line? The value for y at x = 0 changes.

b) Change the slider for the slope. How does it change the regression line? The steepness of the line changes.

c) What happens to the residuals (the red dashed lines) when you change the slope and the intercept of the regression line? The distance between the fitted values (the line) and the observed values (the dots) increases. Therefore, the red dashed lines become longer suggesting that the residuals increase. The model therefore fits the data less well.

4.4.2 Lab activity 2: Statistics anxiety and engagement in module activities

In this lab, we'll be working with real data and using regression to explore the question of whether there is a relationship between statistics anxiety and engagement in course activities. You'll need the data files `psess.csv` and `stars2.csv` and the R-script `122_wk13_labAct2_template.R` that you downloaded when completing Pre-lab activity 3.

4.4.2.1 Background

The hypothesis is that students who are more anxious about statistics are less likely to engage in course-related activities. This avoidance behaviour could ultimately be responsible for lower performance for these students (although we won't be examining the assessment scores in this activity).

We are going to analyse data from the STARS Statistics Anxiety Survey, which was administered to students in the third-year statistics course in Psychology at the University of Glasgow. All the responses have been anonymised by associating the responses for each student with an arbitrary ID number (integer).

The STARS survey (Cruise, Cash, & Bolton, 1985) is a 51-item questionnaire, with each response on a 1 to 5 scale, with higher numbers indicating greater anxiety.

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *Proceedings of the American Statistical Association, Section on Statistical Education*, Las Vegas, NV.

Example items from the STARS survey

Please indicate how much anxiety you would experience (from 1 = no anxiety to 5 = strong anxiety) in each of the following situations:

Studying for an examination in a statistics course

No Anxiety				*Strong Anxiety*
1	2	3	4	5
○	○	○	○	○

Interpreting the meaning of a table in a journal article

No Anxiety				*Strong Anxiety*
1	2	3	4	5
○	○	○	○	○

Going to ask my statistics teacher for individual help with material I am having difficulty understanding

No Anxiety				*Strong Anxiety*
1	2	3	4	5
○	○	○	○	○

Doing the coursework for a statistics course

No Anxiety				*Strong Anxiety*
1	2	3	4	5
○	○	○	○	○

As a measure of engagement in the course, we will use data from Moodle usage analytics. Over the course of the term, there were eight optional weekly on-line sessions that students could attend for extra support. The variable `n_weeks` in the `psess.csv` file tells you how many (out of eight) a given student attended.

Our hypothesis was that greater anxiety would be reflected in lower engagement. Answer the following question.

QUESTION 1: If our hypothesis is correct, what type of correlation (if any) should we observe between students' mean anxiety levels and the variable `n_weeks`?

> **TASK**: Read in both files, have a look at the layout of the data and familiarise yourself with it. **HINT**You can use the `read_csv()` and the `head()` function.

QUESTION 2 In the `stars` table, what do the numbers in the first row across the three columns refer to?

Now that we've read in both data files, the next step is to calculate the mean anxiety scores for each participant. At the moment we have scores on all questions separately for each participant in the `stars` table. Instead we need one mean anxiety score for each participant.

> **TASK**: Write the code to calculate mean anxiety scores. Remember that participant is identified by the `ID` variable. Store the resulting table in a variable named `stars_mean`. **HINT**: Use `group_by()` and `summarise()`. Also, remember to use `na.rm = TRUE` when calculating the mean scores to deal with participants who have missing data (NAs).

QUESTION 3 What is the mean anxiety score for participant 3?

Ok, before we get ahead of ourselves, in order to perform the regression analysis we need to combine the data from `stars` (the mean anxiety scores) with the data from `engage` (n_weeks).

> **TASK**: Join the two tables, call the resulting table `joined`. **HINT**: Use the `inner_join()` function (making use of the variable that is common across both tables) to join.

We now need descriptive statistics for both variables.

> **TASK**: Calculate the mean and standard deviations for the anxiety scores and the engagement data.

QUESTION 4 What are the means and standard deviation for anxiety and engagement with the statistics module?

As always, it is a good idea to visualise your data.Now that we have all the variables in one place, make a scatterplot of anxiety as a function of engagement.

> **TASK**: Write the code to create the scatterplot. **HINT**: For this you'll need the `ggplot()` function together with `geom_point()` and `geom_smooth()`. Make sure to give your axes some sensible labels with the `labs()` function.

QUESTION 5 What does the scatterplot suggest about the relationship between anxiety and engagement?

With all the variables in place, we're ready now to start building the regression model.

> **TASK**: Use the `lm()` function to run the regression model when you model engagement (the outcome variable) as a function of anxiety (the predictor variable) and use the `summary()` function to look at the output. **HINT**: `lm(outcome ~ predictor, data = my_data)`.

QUESTION 6 What is the estimate of the y-intercept for the model, rounded to three decimal places?

QUESTION 7 To three decimal places, if the General Linear Model for this model is `Y=beta0 + beta1X + e`, then `beta1` is …

QUESTION 8 To three decimal places, for each unit increase in anxiety, engagement decreases by …

QUESTION 9 To two decimal places, what is the overall F-ratio of the model?

QUESTION 10 Is the overall model significant?

QUESTION 11 What proportion of the variance does the model explain?

Now that we've fitted a model, let's check whether the model meets the assumptions of linearity, normality and homoscedasticity.

> **TASK**: Write the code to check the assumptions. **HINT**: `crPlots()` to check linearity, `qqPlot()` to check normality of the residuals, and `residualPlot()` to check homoscedasticity of the residuals.

QUESTION 12 Does the relationship appear to be linear?

QUESTION 13 Do the residuals show normality?

QUESTION 14 Do the residuals show homoscedasticity?

Finally, it's time to write up the results following APA guidelines.

QUESTION 15 What would the results section look like if you wrote them up, following APA guidelines? **HINT**: The **Purdue writing lab website** is helpful for guidance on punctuating statistics.

4.5 Answers

When you have completed all of the lab content, you may want to check your answers with our completed version of the script for this week. **Remember**, looking at this script (studying/revising it) does not replace the process of working through the lab activities, trying them out for yourself, getting stuck, asking questions, finding solutions, adding your own comments, etc. **Actively engaging** with the material is the way to learn these analysis skills, not by looking at someone else's completed code…

The answers to the questions and the script containing the code will be available after the final lab session has taken place.