Welcome

Professor Andrew Stewart, Twitter: @ajstewart_lang

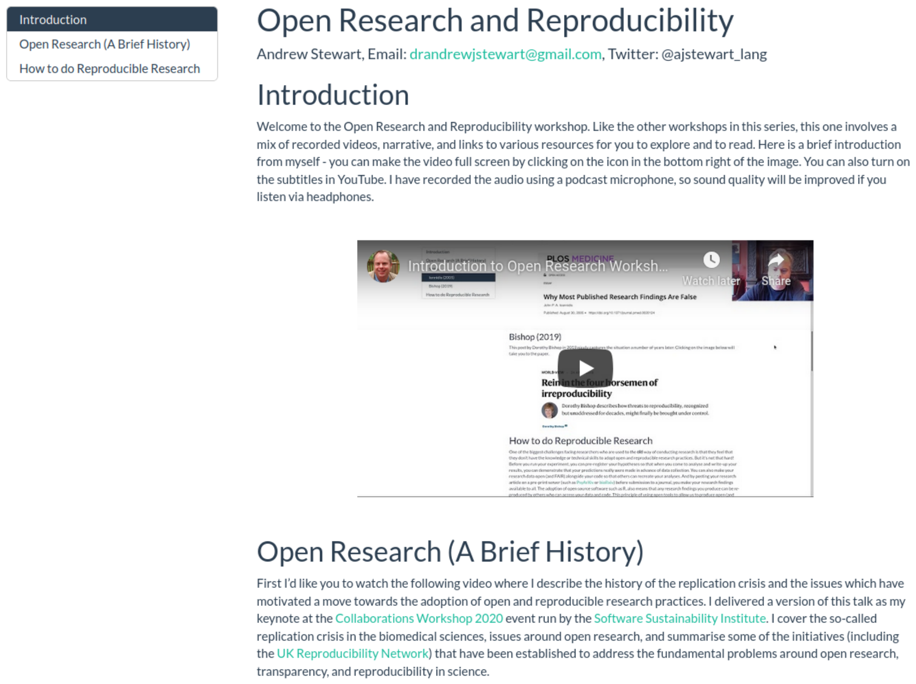
Contents

Welcome to the Advanced Data Skills, Open Science and Reproducibility M.Res. unit BIOL63101. The following video (as with all the others in this unit) is best viewed in fullscreen mode. I have recorded the audio with a podcasting microphone, so it's best listened to with headphones. YouTube generates subtitles automatically, so please turn those in if you'd find them useful.

Workshop 1

Open Research and Reproducibility

In this workshop I will first introduce you to the key concepts in open research, and talk about the so-called "replication crisis" in the Psychological, Biomedical, and Life Sciences that has resulted in the Open Research movement I will also discuss the importance of adopting reproducible research practices in your own research, and provide an introduction to various tools and processes you can incorporate into your own research workflows that will allow you to conduct reproducible research. To go to the first part of the workshop, just click on the image below.


Introduction
Open Research (A Brief History)
How to do Reproducible Research
Introduction to Open Research Worksh...
Watch later
Share
PLOS MEDICINE
Why Most Published Research Findings Are False
Bishop (2019)
Rein in the four horsemen of irreproducibility
How to do Reproducible Research

Open Research and Reproducibility

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Introduction

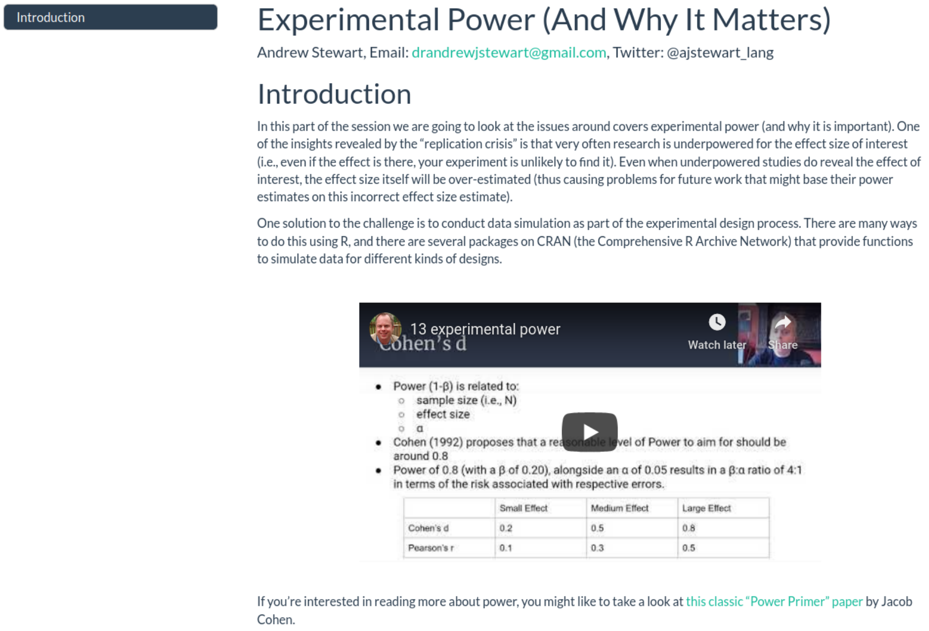
Welcome to the Open Research and Reproducibility workshop. Like the other workshops in this series, this one involves a mix of recorded videos, narrative, and links to various resources for you to explore and to read. Here is a brief introduction from myself - you can make the video full screen by clicking on the icon in the bottom right of the image. You can also turn on the subtitles in YouTube. I have recorded the audio using a podcast microphone, so sound quality will be improved if you listen via headphones.

Open Research (A Brief History)

First I'd like you to watch the following video where I describe the history of the replication crisis and the issues which have motivated a move towards the adoption of open and reproducible research practices. I delivered a version of this talk as my keynote at the Collaborations Workshop 2020 event run by the Software Sustainability Institute. I cover the so-called replication crisis in the biomedical sciences, issues around open research, and summarise some of the initiatives (including the UK Reproducibility Network) that have been established to address the fundamental problems around open research, transparency, and reproducibility in science.

Experimental Power

The second part of this workshop covers experimental power (and why it is important). One of the insights revealed by the "replication crisis" is that very often research is underpowered for the effect size of interest (i.e., even if the effect is there, your experiment is unlikely to find it). Many of the issues stem from researchers not spending sufficient time considering the power aspects of their research design. In this workshop, I'll provide you with an overview of some of the issues - just click on the image below to view this second part of the workshop:



Experimental Power (And Why It Matters)

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Introduction

In this part of the session we are going to look at the issues around covers experimental power (and why it is important). One of the insights revealed by the "replication crisis" is that very often research is underpowered for the effect size of interest (i.e., even if the effect is there, your experiment is unlikely to find it). Even when underpowered studies do reveal the effect of interest, the effect size itself will be over-estimated (thus causing problems for future work that might base their power estimates on this incorrect effect size estimate).

One solution to the challenge is to conduct data simulation as part of the experimental design process. There are many ways to do this using R, and there are several packages on CRAN (the Comprehensive R Archive Network) that provide functions to simulate data for different kinds of designs.

Cohen's d

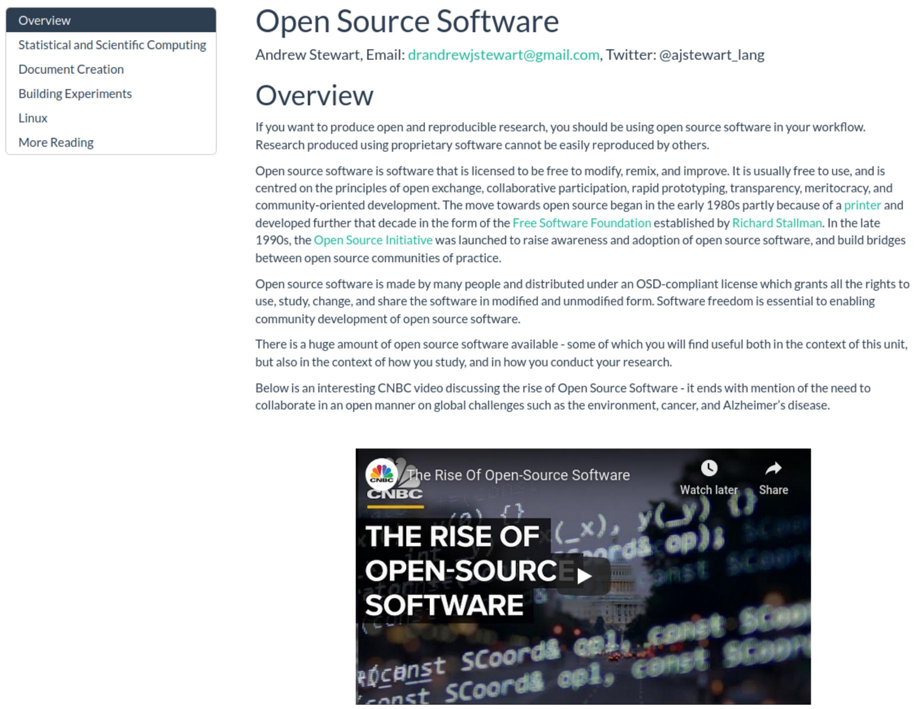
- Power (1-β) is related to:
  - sample size (i.e., N)
  - effect size
  - α
- Cohen (1992) proposes that a reasonable level of Power to aim for should be around 0.8
- Power of 0.8 (with a β of 0.20), alongside an α of 0.05 results in a β:α ratio of 4:1 in terms of the risk associated with respective errors.

	Small Effect	Medium Effect	Large Effect
Cohen's d	0.2	0.5	0.8
Pearson's r	0.1	0.3	0.5

If you're interested in reading more about power, you might like to take a look at this classic "Power Primer" paper by Jacob Cohen.

Open Source Software

The third part of the workshop involves a very brief overview of open source software, the use of which is arguably key for researchers to be able to adopt open and reproducible research workflows. To view this third part, just click on the image below.



Open Source Software

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

If you want to produce open and reproducible research, you should be using open source software in your workflow. Research produced using proprietary software cannot be easily reproduced by others.

Open source software is software that is licensed to be free to modify, remix, and improve. It is usually free to use, and is centred on the principles of open exchange, collaborative participation, rapid prototyping, transparency, meritocracy, and community-oriented development. The move towards open source began in the early 1980s partly because of a printer and developed further that decade in the form of the Free Software Foundation established by Richard Stallman. In the late 1990s, the Open Source Initiative was launched to raise awareness and adoption of open source software, and build bridges between open source communities of practice.

Open source software is made by many people and distributed under an OSD-compliant license which grants all the rights to use, study, change, and share the software in modified and unmodified form. Software freedom is essential to enabling community development of open source software.

There is a huge amount of open source software available - some of which you will find useful both in the context of this unit, but also in the context of how you study, and in how you conduct your research.

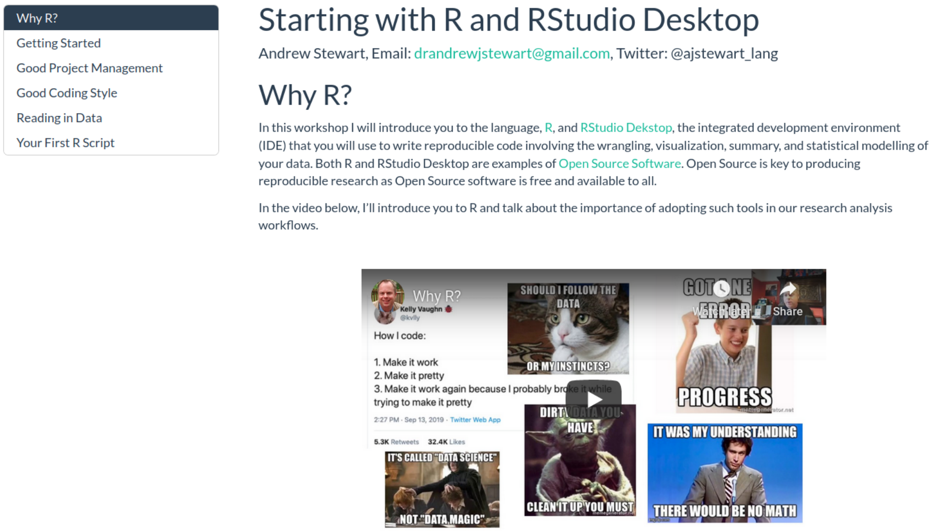
Below is an interesting CNBC video discussing the rise of Open Source Software - it ends with mention of the need to collaborate in an open manner on global challenges such as the environment, cancer, and Alzheimer's disease.

CNBC
The Rise Of Open-Source Software
Watch later
Share
THE RISE OF OPEN-SOURCE SOFTWARE

Workshop 2

Starting with R and RStudio Desktop

In this workshop I'll introduce you to R (the language) and RStudio Desktop (the environment we use to interact with the language). I've also added a link to a great talk by the founder of RStudio, J.J. Alaire. At the end of the workshop I have put together a video which will show you how to run your first R script.



Starting with R and RStudio Desktop

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Why R?

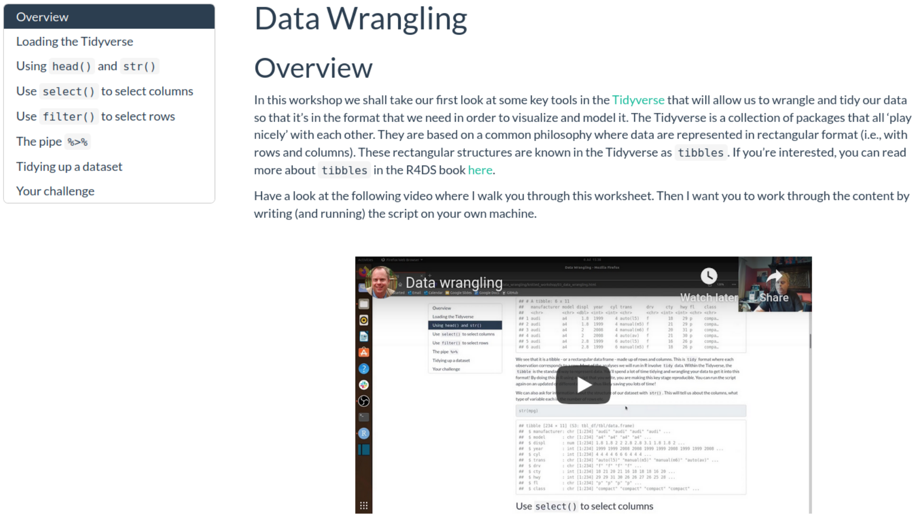
In this workshop I will introduce you to the language, R, and RStudio Dekstop, the integrated development environment (IDE) that you will use to write reproducible code involving the wrangling, visualization, summary, and statistical modelling of your data. Both R and RStudio Desktop are examples of Open Source Software. Open Source is key to producing reproducible research as Open Source software is free and available to all.

In the video below, I'll introduce you to R and talk about the importance of adopting such tools in our research analysis workflows.

Workshop 3

Data Wrangling

In this workshop I will introduce you to a number of key packages known as the `Tidyverse` These packages contain a large number of functions for working with data in tidy format. By making our data wrangling reproducible (i.e., by coding it in R), we can easily re-run this stage of our analysis pipeline as new data gets added. Reproducibility of the data wrangling stage is a key part of the analysis process and often gets overlooked in terms of needing to ensure it is reproducible. There are two parts to this workshop. The first focuses on data wrangling/tidying. To go to this first part, just click on the image below.



- Overview
- Loading the Tidyverse
- Using `head()` and `str()`
- Use `select()` to select columns
- Use `filter()` to select rows
- The pipe `%>%`
- Tidying up a dataset
- Your challenge

Data Wrangling

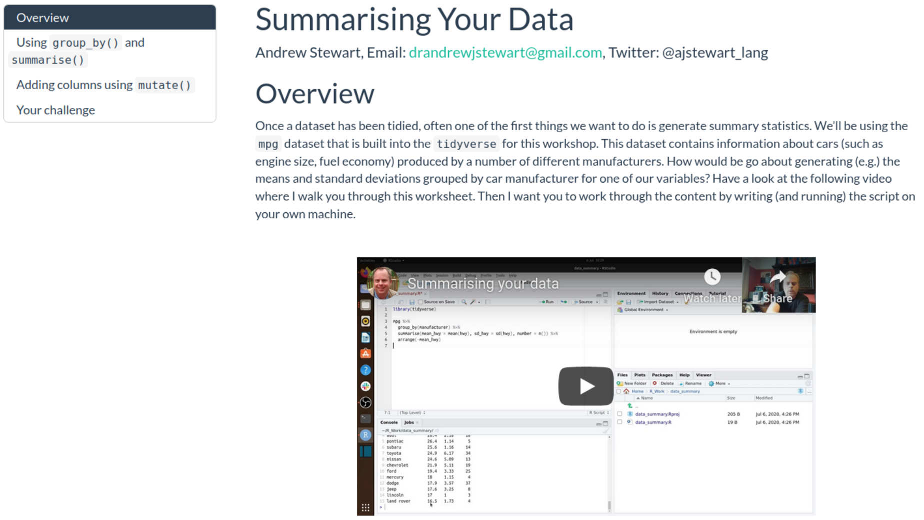
Overview

In this workshop we shall take our first look at some key tools in the Tidyverse that will allow us to wrangle and tidy our data so that it's in the format that we need in order to visualize and model it. The Tidyverse is a collection of packages that all 'play nicely' with each other. They are based on a common philosophy where data are represented in rectangular format (i.e., with rows and columns). These rectangular structures are known in the Tidyverse as `tibbles`. If you're interested, you can read more about `tibbles` in the R4DS book here.

Have a look at the following video where I walk you through this worksheet. Then I want you to work through the content by writing (and running) the script on your own machine.

Summarising Your Data

Once you have completed this first part and have your R script up and running, click on the image below for the second part where you'll learn how to aggregate and summarise your data.



Summarising Your Data

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

Once a dataset has been tidied, often one of the first things we want to do is generate summary statistics. We'll be using the `mpg` dataset that is built into the `tidyverse` for this workshop. This dataset contains information about cars (such as engine size, fuel economy) produced by a number of different manufacturers. How would be go about generating (e.g.) the means and standard deviations grouped by car manufacturer for one of our variables? Have a look at the following video where I walk you through this worksheet. Then I want you to work through the content by writing (and running) the script on your own machine.

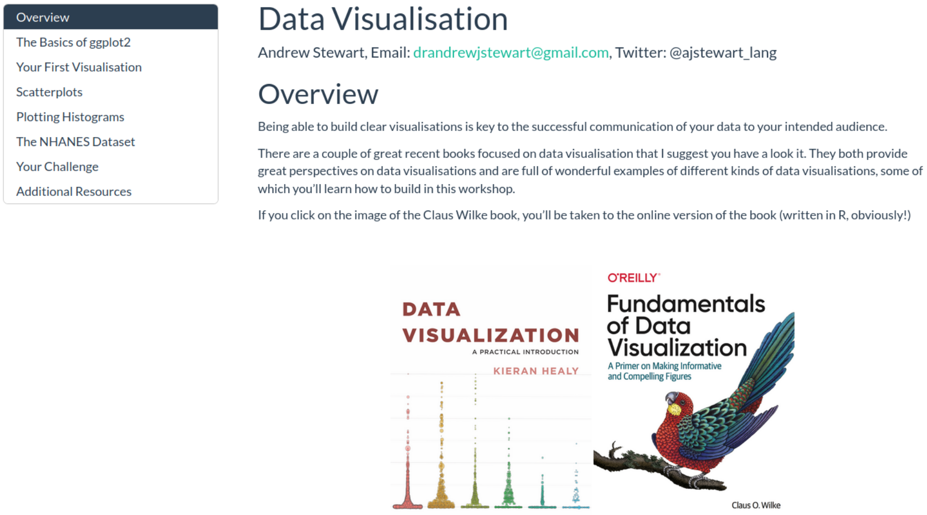
```
library(tidyverse)

mpg %>%
  group_by(manufacturer) %>%
  summarise(mean_hwy = mean(hwy), sd_hwy = sd(hwy), number = n()) %>%
  arrange(-mean_hwy)
```

```
 5 pontiac       26.4   1.14      5
 6 subaru        25.6   1.16     14
 7 toyota        24.9   6.17     34
 8 nissan        24.6   5.09     13
 9 chevrolet     21.9   5.11     19
10 ford          19.4   3.33     25
11 mercury       18     1.15      4
12 dodge         17.9   3.57     37
13 jeep          17.6   3.25      8
14 lincoln       17     1         3
15 land rover    16.5   1.73      4
>
```

Workshop 4

Data Visualisation

In this workshop we will explore the basics of Data Visualization using R. You'll have the opportunity to write an R script on your own computer that will generate some nice data visualisations. Just click on the image below to start.



Data Visualisation

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

Being able to build clear visualisations is key to the successful communication of your data to your intended audience.

There are a couple of great recent books focused on data visualisation that I suggest you have a look it. They both provide great perspectives on data visualisations and are full of wonderful examples of different kinds of data visualisations, some of which you'll learn how to build in this workshop.

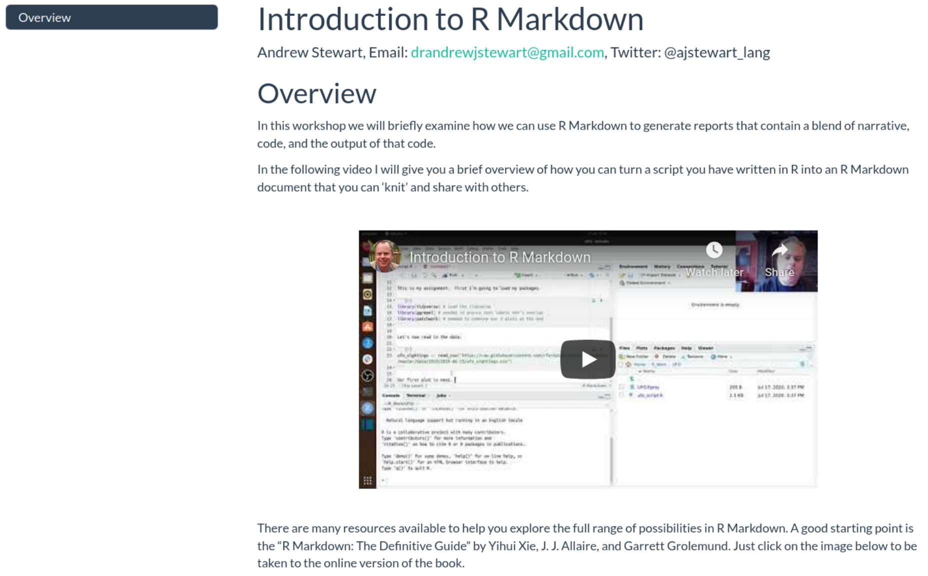
If you click on the image of the Claus Wilke book, you'll be taken to the online version of the book (written in R, obviously!)

DATA VISUALIZATION
A PRACTICAL INTRODUCTION
KIERAN HEALY
O'REILLY
Fundamentals of Data Visualization
A Primer on Making Informative and Compelling Figures
Claus O. Wilke

Workshop 5

R Markdown

In this workshop I will show you how to generate a report in `.html` format using R Markdown. Reports written using R Markdown allow you to combine narrative that you've written alongwith R code chunks, and the output associated with those code chunks all in one `knitted` document. The assignments for this unit need to be produced using R Markdown.



Introduction to R Markdown

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

In this workshop we will briefly examine how we can use R Markdown to generate reports that contain a blend of narrative, code, and the output of that code.

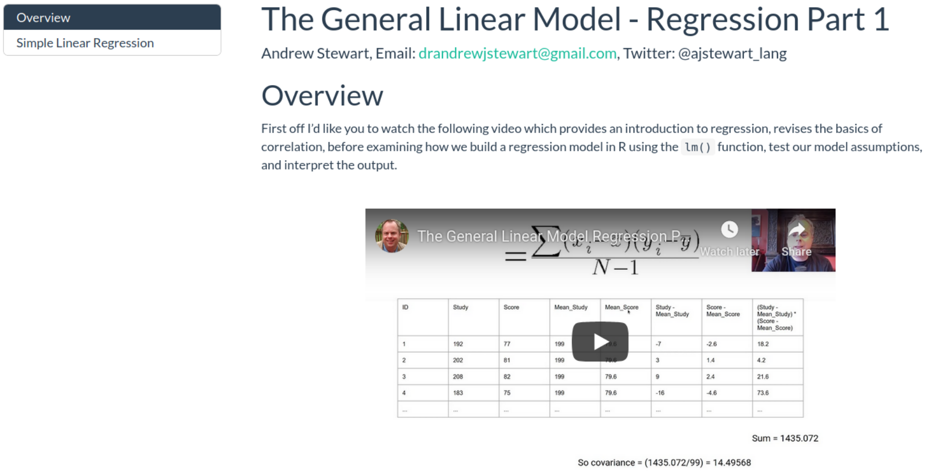
In the following video I will give you a brief overview of how you can turn a script you have written in R into an R Markdown document that you can 'knit' and share with others.

There are many resources available to help you explore the full range of possibilities in R Markdown. A good starting point is the "R Markdown: The Definitive Guide" by Yihui Xie, J. J. Allaire, and Garrett Grolemund. Just click on the image below to be taken to the online version of the book.

Workshop 6

Regression Part 1

In this workshop we will explore Simple Regression in the context of the General Linear Model (GLM). You will also have the opportunity to build some regression models where you predict an outcome variable on the basis of one predictor. You will also learn how to run model diagnostics to ensure you are not violating any key assumptions of regression.



The General Linear Model - Regression Part 1

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

First off I'd like you to watch the following video which provides an introduction to regression, revises the basics of correlation, before examining how we build a regression model in R using the `lm()` function, test our model assumptions, and interpret the output.

$$= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

ID	Study	Score	Mean_Study	Mean_Score	Study - Mean_Study	Score - Mean_Score	(Study - Mean_Study) * (Score - Mean_Score)
1	192	77	199	79.6	-7	-2.6	18.2
2	202	81	199	79.6	3	1.4	4.2
3	208	82	199	79.6	9	2.4	21.6
4	183	75	199	79.6	-16	-4.6	73.6
...	...	...	...	...	...	...	...

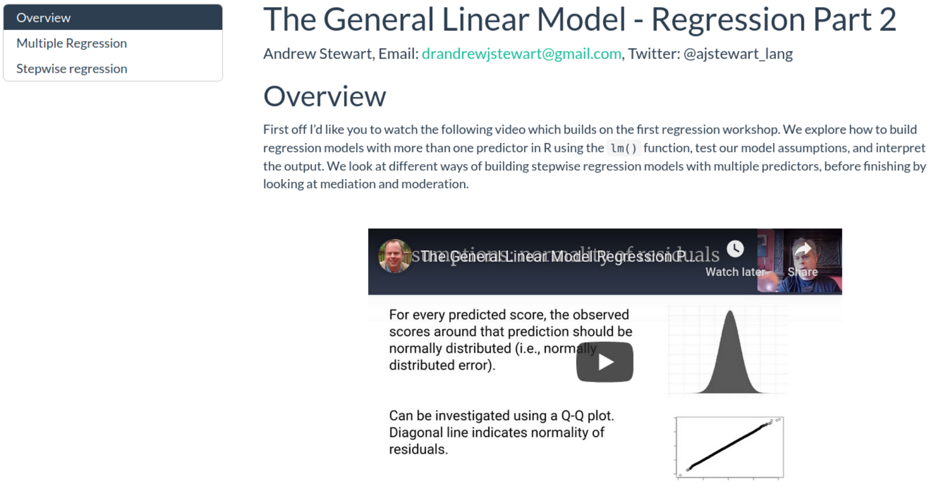
Sum = 1435.072

So covariance = (1435.072/99) = 14.49568

Workshop 7

Regression Part 2

In this workshop will explore Multiple Regression in the context of the General Linear Model (GLM). Multiple Regressions builds on Simple Regression, except that having one predictor (as is the case with Simple Regression) we will be dealing with multiple predictors. Again, you will have the opporunity to build some regression models and use various methods to decide which one is 'best'. You will also learn how to run model disagnostics for these models as you did in the case of Simple Regression.



The General Linear Model - Regression Part 2

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

First off I'd like you to watch the following video which builds on the first regression workshop. We explore how to build regression models with more than one predictor in R using the `lm()` function, test our model assumptions, and interpret the output. We look at different ways of building stepwise regression models with multiple predictors, before finishing by looking at mediation and moderation.

The General Linear Model Regression P...
sumptions: normality of residuals
Watch later
Share

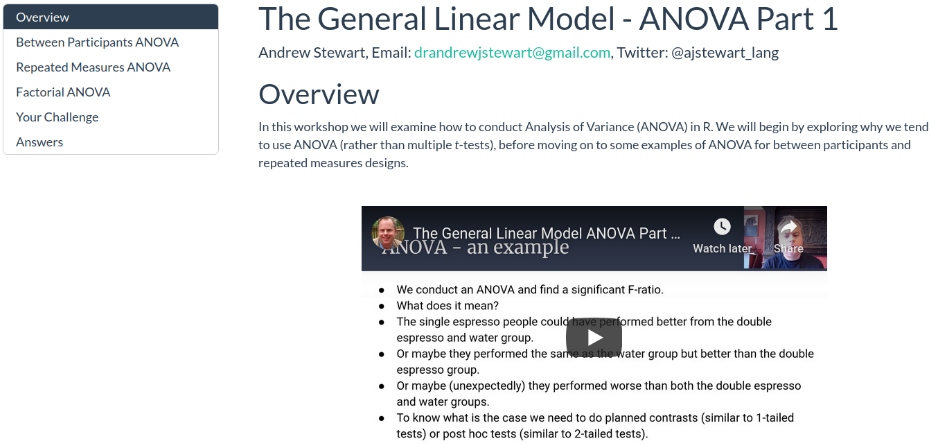
For every predicted score, the observed scores around that prediction should be normally distributed (i.e., normally distributed error).

Can be investigated using a Q-Q plot. Diagonal line indicates normality of residuals.

Workshop 8

ANOVA Part 1

In this workshop we will explore Analysis of Variance (ANOVA) in the context of model building in R for between participants designs, repeated measures designs, and factorial designs. You will learn how to use the {afex} package for building models with Type III Sums of Squares, and the {emmeans} package to conduct follow up tests to explore main effects and interactions.



The General Linear Model - ANOVA Part 1

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

In this workshop we will examine how to conduct Analysis of Variance (ANOVA) in R. We will begin by exploring why we tend to use ANOVA (rather than multiple *t*-tests), before moving on to some examples of ANOVA for between participants and repeated measures designs.

The General Linear Model ANOVA Part ...

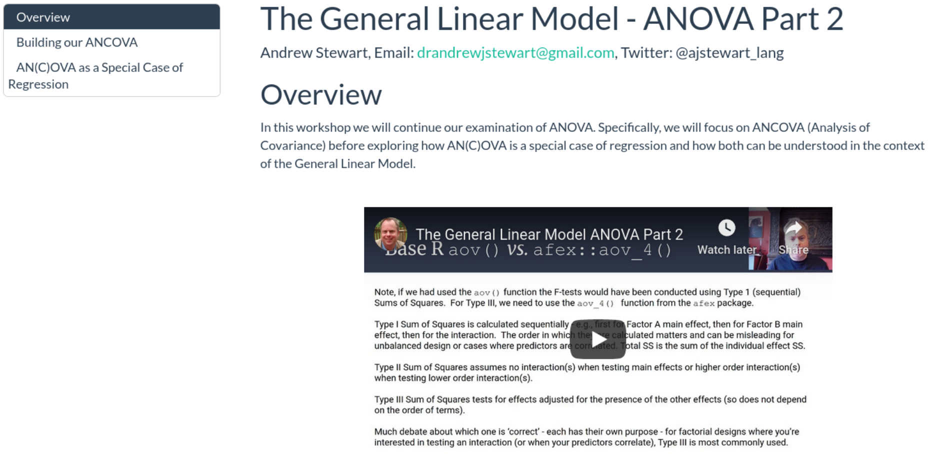
ANOVA – an example

- We conduct an ANOVA and find a significant F-ratio.
- What does it mean?
- The single espresso people could have performed better from the double espresso and water group.
- Or maybe they performed the same as the water group but better than the double espresso group.
- Or maybe (unexpectedly) they performed worse than both the double espresso and water groups.
- To know what is the case we need to do planned contrasts (similar to 1-tailed tests) or post hoc tests (similar to 2-tailed tests).

Workshop 9

ANOVA Part 2

In this workshop we will build on Workshop 8 to explore Analysis of Covariance (ANCOVA). In this workshop we will also examine ANOVA and ANCOVA as special cases of regression and see how we can build both via a linear model. By then doing this yourselves, you wil hopefully be convinced that ANOVA and regression are really the same thing.



The General Linear Model - ANOVA Part 2

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

In this workshop we will continue our examination of ANOVA. Specifically, we will focus on ANCOVA (Analysis of Covariance) before exploring how AN(C)OVA is a special case of regression and how both can be understood in the context of the General Linear Model.

Base R `aov()` *vs.* `afex::aov_4()`

Note, if we had used the `aov()` function the F-tests would have been conducted using Type 1 (sequential) Sums of Squares. For Type III, we need to use the `aov_4()` function from the `afex` package.

Type I Sum of Squares is calculated sequentially - e.g., first for Factor A main effect, then for Factor B main effect, then for the interaction. The order in which they are calculated matters and can be misleading for unbalanced design or cases where predictors are correlated. Total SS is the sum of the individual effect SS.

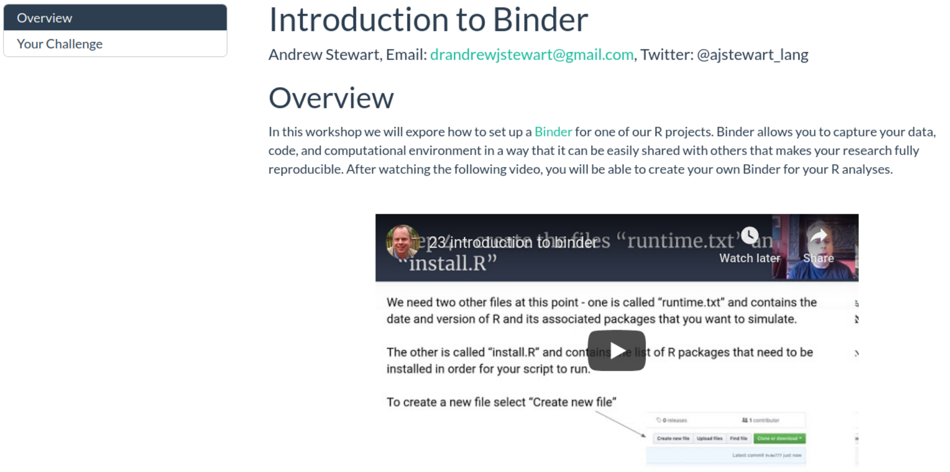
Type II Sum of Squares assumes no interaction(s) when testing main effects or higher order interaction(s) when testing lower order interaction(s).

Type III Sum of Squares tests for effects adjusted for the presence of the other effects (so does not depend on the order of terms).

Much debate about which one is 'correct' - each has their own purpose - for factorial designs where you're interested in testing an interaction (or when your predictors correlate), Type III is most commonly used.

Workshop 10

This workshop will introduce you to Binder, a tool for allowing you to reproduce your computational environment, as well as your code and data, and share all of these with a collaborator via sharing a simple web link.



Introduction to Binder

Andrew Stewart, Email: drandrewjstewart@gmail.com, Twitter: @ajstewart_lang

Overview

In this workshop we will expore how to set up a Binder for one of our R projects. Binder allows you to capture your data, code, and computational environment in a way that it can be easily shared with others that makes your research fully reproducible. After watching the following video, you will be able to create your own Binder for your R analyses.

ep 4 – create the files "runtime.txt" an "install.R"

We need two other files at this point - one is called "runtime.txt" and contains the date and version of R and its associated packages that you want to simulate.

The other is called "install.R" and contains the list of R packages that need to be installed in order for your script to run.

To create a new file select "Create new file"

Technical Details

The structure for this unit was very much inspired by the Sharing At Short Notice webinar by Alison Hill and Desirée De Leon.

The repo for each workshop can be accessed via the 'Improve this Workshop' link at the bottom of each workshop page. The workshops and this website were all written using R Markdown and the website is hosted on Netlify via continunous deployment from this GitHub repository.