Basic Statistical Models Using R

Andrew Stewart and Peter Smyth

@ajstewart_lang

BETTER SOFTWARE BETTER RESEARCH

www.software.ac.uk

Software Sustainability Institute

The Challenges

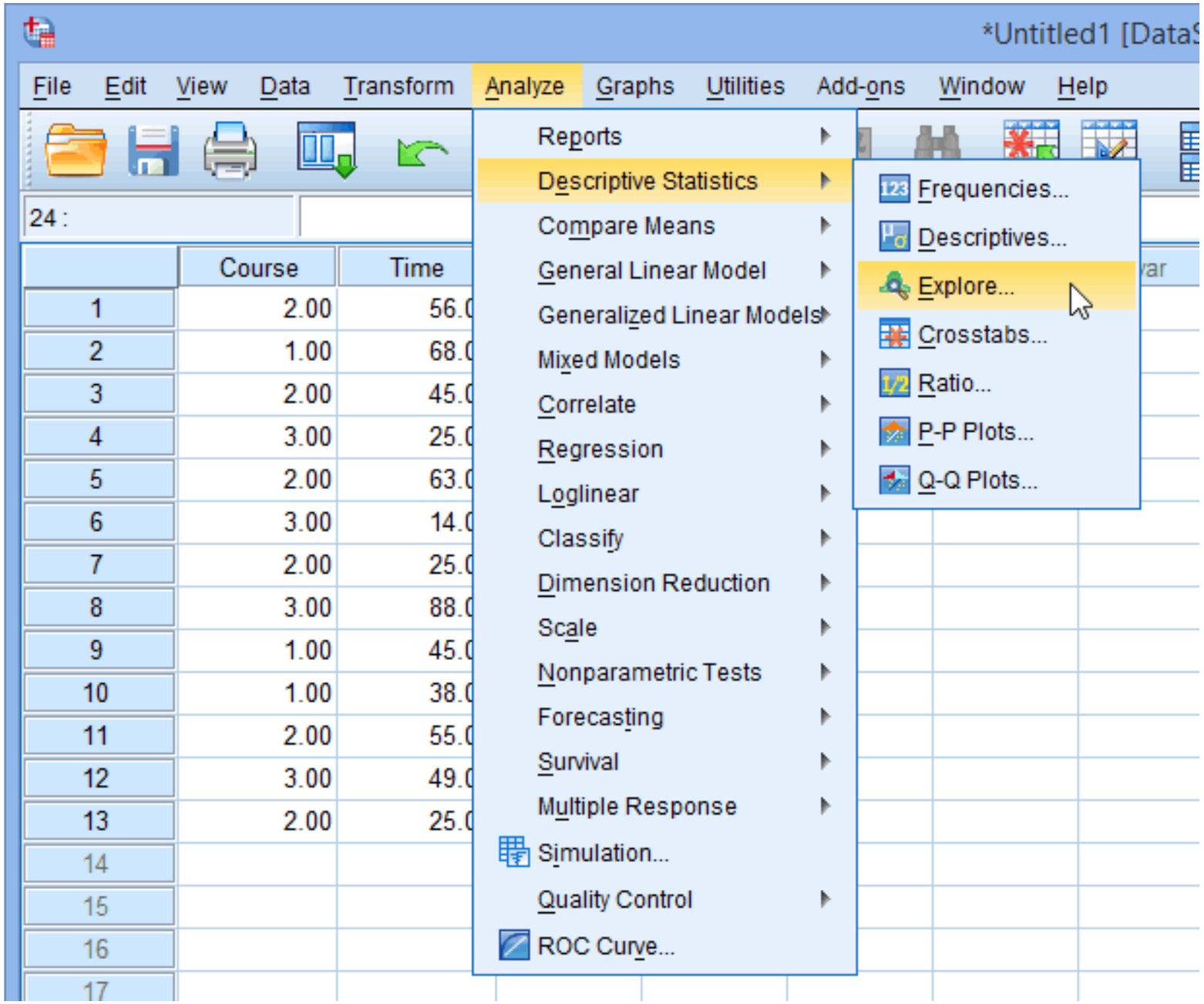
- Teaching Statistics and R to M-level students in Psychology - 2 units on MRes and new MSci unit.

- Mixed background and interest in Statistics.

- Mixed background and interest in coding (with many never having coded before).

- Most people's UG backgrounds are similar…

Pointy and Clicky

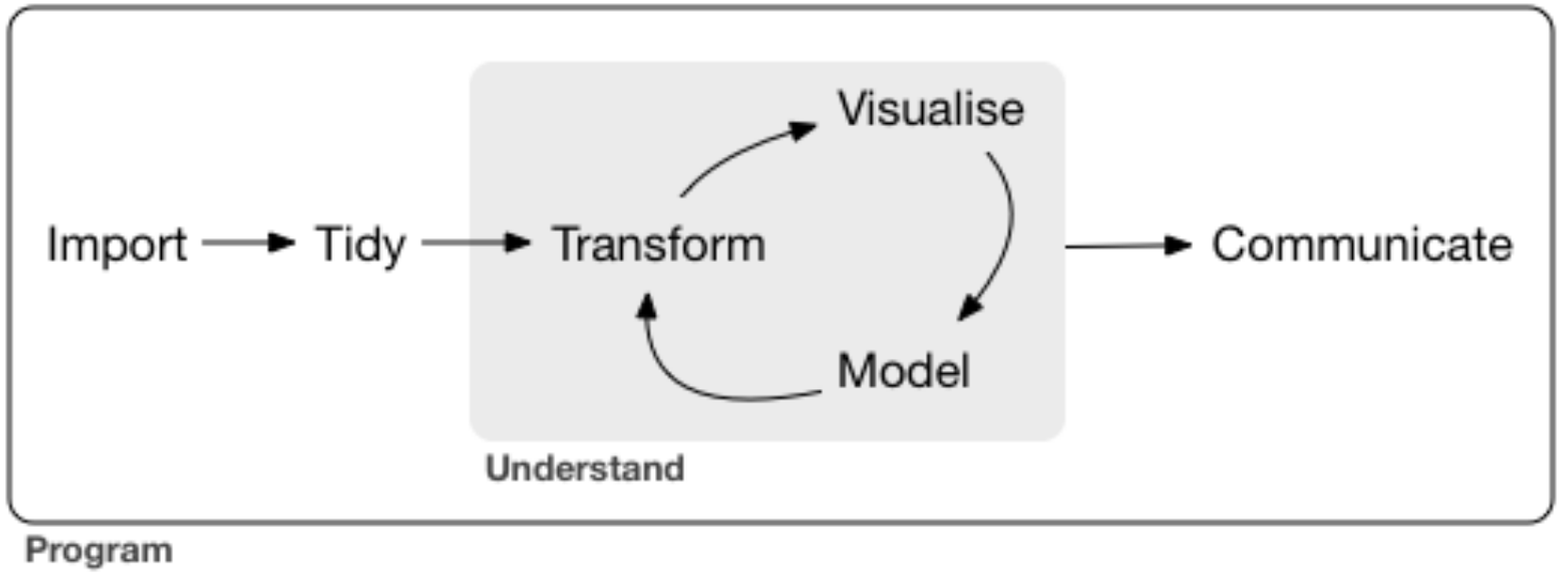


The problem is that you can't really do reproducible analysis in a GUI…

And it's not so much as statistical knowledge, but knowledge about what to click (and where)…

- Reproducibility is key in science - and one of the easiest ways to engage in reproducible research is to use an open source statistical language such as R.

- The tidyverse workflow allows for data importing, wrangling, visualisation, and modelling all in the same reproducible workflow.



Hadley Wickham and Garrett Grolemund

What we'll cover in this session…

- Data simulation
- Data visualisation
- t-tests
- General linear model (continuous predictors)
- General linear model (factorial)

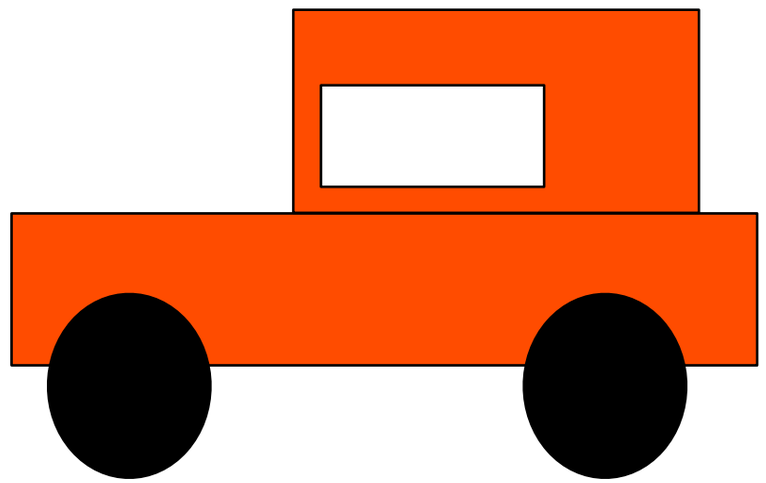
What packages we'll use in this session…

```
install.packages("tidyverse")
install.packages("broom")
install.packages("afex")
install.packages("emmeans")

library(tidyverse)
library(broom)
library(afex)
library(emmeans)
```

Testing for differences between groups using Student's t-distribution.

Real data



Model 1

Model 2

- So how do we tell if a particular statistical model is a good fit to our data?
- We can look at how well a model fits our data.

Real data

=

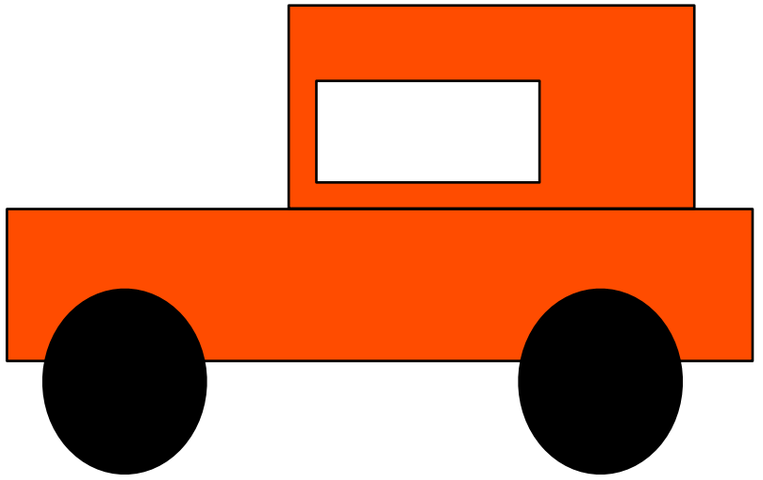
Model 1

+ Error

Real data

=

Model 2



+ Error

- We want to select the model which fits the data best (e.g., has the smallest 'errror').

Let’s start coding…

.Rmd file

https://bit.ly/2YaYg9V

Markdown

https://bit.ly/2LcmP2D

Multiple Regression

- Multiple regression is just an extension of simple linear regression - but rather than having one predictor we have several.

- Our goal is the same though - to end up with an equation for a straight line that is the best fit to our data - in other words, we want to minimise the residual error.

Multicollinearity and Singularity

- Multicollinearity: when two or more variables are highly correlated (tested by examining VIF value for each variable).

- Singularity: redundancy (e.g., one of the variables is a combination of two or more of the other IVs).

- We can use collinearity diagnostics to see if we have a possible problem…

Assumptions: no multicollinearity among predictors

- VIF stands for Variance Inflation Factor. Essentially, it tells us about when we have to worry about (multi)collinearity. We can ask for VIF to any model in R by using the function `vif()` in the `car` package.

- So when do you worry?

- As a rule of thumb VIF greater than 10 suggests a multicollinearity issue (although greater than 5 has been suggested too - more conservative).

Multiple Regression

.Rmd file

https://bit.ly/2J4BDOa

Markdown

https://bit.ly/2J1glvs