Getting Started and R Markdown

STAT3022 Applied Linear Models Lecture 1

2020/02/24

THE UNIVERSITY OF SYDNEY

Today

1. Introduce yourself to someone you don't know.
2. Course structure and expectations.
3. Meet the toolkit - learn to use R Markdown.

Disability Services Support

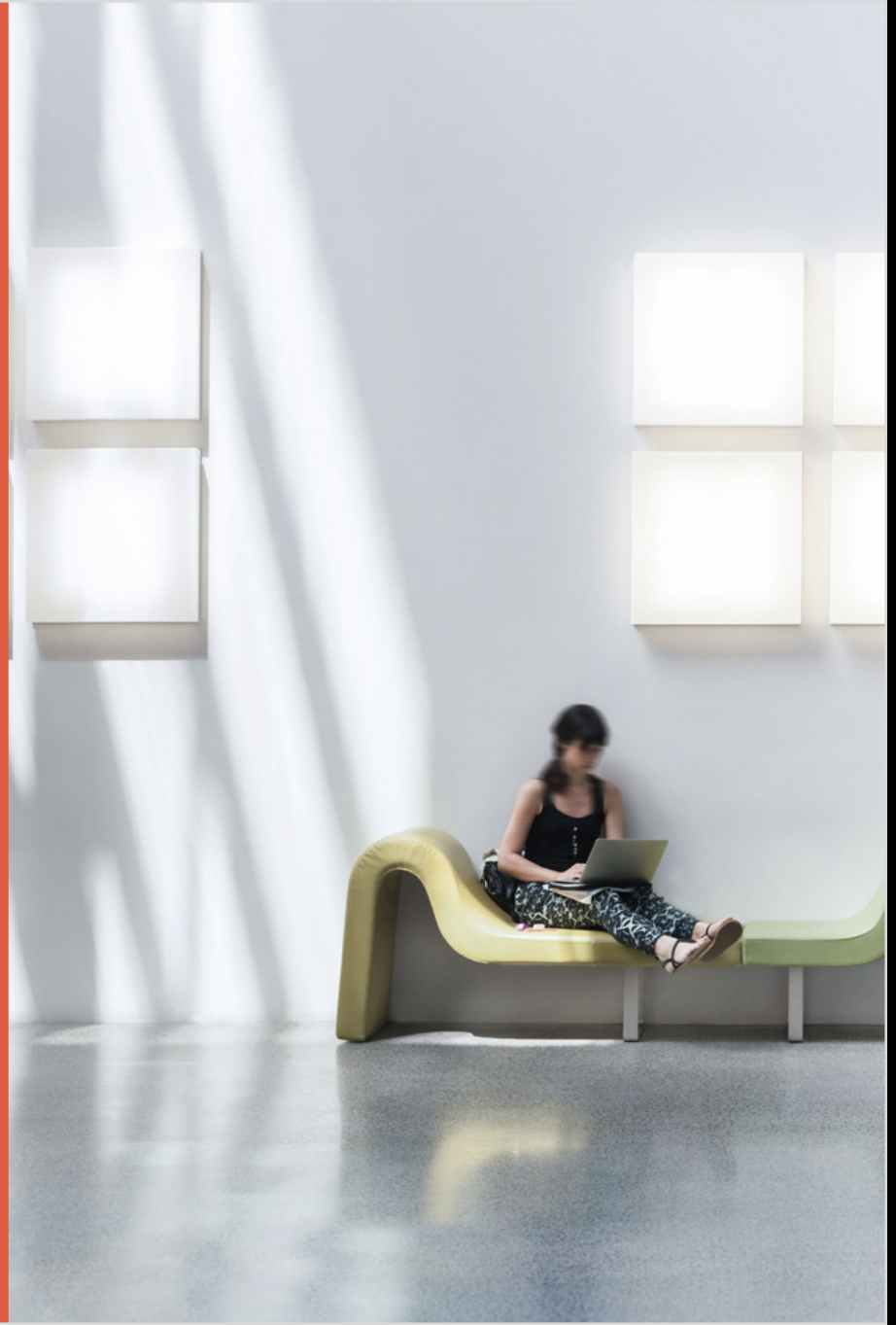
You may not think of yourself as having a disability but the definition under the **Disability Discrimination Act (1992)** is broad and includes temporary or chronic medical conditions, physical or sensory disabilities, psychological conditions and learning disabilities.

Some types of disabilities we see include:
Anxiety // Arthritis // Asthma // Autism // ADHD
Bipolar disorder // Broken bones // Cancer
Cerebral palsy // Chronic fatigue syndrome
Crohn's disease // Cystic fibrosis // Depression
Diabetes // Dyslexia // Epilepsy // Hearing impairment // Learning disability // Mobility impairment // Multiple sclerosis // Post-traumatic stress // Schizophrenia // Vision impairment

Students needing assistance are advised to register with Disability Services as early as possible. Please contact us or review our website to find out more.

THE UNIVERSITY OF SYDNEY

Disability Services Office
sydney.edu.au/disability
02 8627 8422



Lecturers

- A/Prof. Jennifer Chan for STAT3022
  - jennifer.chan@sydney.edu.au
  - Carslaw 817
- Dr. Munir Hiabu for STAT3922
  - munir.hiabu@sydney.edu.au
  - Carslaw 827

STAT3022

What is this unit about?

The overall aim of this unit is to develop skills in the **statistical analysis** of data from **designed experiments** and **observational studies**.

The unit will be divided into five themes:

STAT3022 Course Structure

Structure

- 3 one hour lectures per week on Mon, Thu & Fri 9am from week 1 to 13
- 1 tutorial per week from week 2 to 13
- 1 computer lab per week from week 1 to 13 (replaced by lecture for STAT3922 students)

Assessments

- 2 assignments due Mon 6th April and Mon 25th May
  - each worth 7.5% of total mark for STAT3022 students
  - each worth 5% of total mark for STAT3922 students
- 1 quiz held 9am Thu 7th May
  - each worth 15% of total mark for STAT3022 students
  - each worth 12.5% of total mark for STAT3922 students
- final exam worth 55% of total mark during June exam period

Academic Integrity

Assignments

- You may make use of any online resources but you must explicitly cite where you obtained any information you directly use (or use as inspiration).
- Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.
- On assignments you may not directly share code with another student in this class.
- You are welcome to discuss the problems together and ask for advice, but you may not send or make use of code from another person.

Sourced from Data Science in a Box: https://datasciencebox.org/

? Help!

Where to get help

- Post your questions on edstem.
- Ask your tutor during tutorial or lab or lecturer during consultation hour: 10-11am Fridays.

Tips for asking questions

- *First search* existing resources and discussion for answers.
- Use proper formatting in edstem. E.g. if using code, use code formatting and LaTeX for mathematical equations.
- Give context and be precise in your description:
  - Good description: "For Tutorial 3 Q1 (k), the answer is given as $50.59165 \pm 1.812461 \times 4.567673$. Where does the value 1.812461 come from? Should this be $t_{10}^{1}(0.975) = 2.228$, which is the same as question 1 (j)?"
  - Bad description: "I don't get Tutorial 3 Q1 (k).

Expectations

Organisation

You are expected to:

- check the STAT3022 canvas website frequently;
- check and *contribute* to the STAT3022 edstem discussion board;
- complete the tutorial and computer lab questions in Week n by Week $n+1$;
- seek help from your tutor during tutorial and computer lab; and
- post your questions in edstem discussion board.

Technical skills

You are expected to:

- program in statistical programming language R; if you are not familiar with R, you are recommended to check the resources page in canvas and quickly catch up with the basics in R; and
- completed (STAT2x12 or DATA2x02) and MATH1x02 or equivalent courses.

What's your expectations out of this course?

Theme 2 Overview

Linear regression

- Model fit
  - Simple & Multiple linear regression
  - Polynomial regression
  - Robust regression
- Model diagnostics
  - Leverage points and outliers
  - Multi-collinearity
  - Goodness of fit measures: Multiple correlation coefficient, AIC, Cp and BIC
- Model selection
  - Forward, backwards, step-wise selection procedures
  - Inference: t-test and general F-test

Did someone say "Applied"?

Theme 1

Toolkits for analysis

R

+

RStudio

+

tidyverse ggplot2 dplyr rmarkdown ...

What's the difference between these?

R essentials

A short list:

- Functions are (most often) verbs, followed by what they will be applied to in parantheses:

```
do_this(to_this)
do_that(to_this, to_that, with_those)
```

- Columns (variables) in data frames are accessed with `$`:

```
dataframe$var_name
```

- Packages are installed with the install.packages function and loaded with the library function, once per session:

```
install.packages("package_name")
library(package_name)
```

Sourced from Data Science in a Box: https://datasciencebox.org/

Literate Programming

In literature programming, you interweave code, output and narrative in one place.

Reproducibility checklist

What does it mean for a data analysis to be "reproducible"?

Near-term goals:

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear why it was done? (e.g., how were parameter settings chosen?)

Long-term goals:

- Can the code be used for other data?
- Can you extend the code to do other things?

Side note: Literate Programming by Donald Knuth is the seminal book on literate programming.

Sourced from Data Science in a Box: https://datasciencebox.org/

Markdown Syntax

- You will achieve literate programming by using R Markdown in this course.
- Let's start with simple markdown syntax for text.

Emphasis

```
*italic*   **bold**

_italic_   __bold__
```

Headers

```
# Header 1

## Header 2

### Header 3
```

Lists

Unordered List

```
* Item 1
* Item 2
    + Item 2a
    + Item 2b
```

Ordered List

```
1. Item 1
2. Item 2
3. Item 3
    + Item 3a
    + Item 3b
```

Manual Line Breaks

End a line with two or more spaces:

```
Roses are red,
Violets are blue.
```

Links

Use a plain http address or add a link to a phrase:

```
http://example.com

[linked phrase](http://example.com)
```

Images

Images on the web or local files in the same directory:

```
![alt text](http://example.com/logo.png)

![alt text](figures/img.png)
```

Blockquotes

```
A friend once said:

> It's always better to give
> than to receive.
```

Markdown Example

Look here

See also RStudio >

Help >

Cheatsheet >

R Markdown Reference Guide

Code

Headers

```
# First level header
## Second level header
### Third level header
```

Emphasis

```
*This text will be italic*
**This text will be bold**
*You **can** combine them*
```

Images

```
![](images/kunoichi.svg)
```

Output

First level header

Second level header

Third level header

This text will be italic

This text will be bold

*You **can** combine them*

An R Markdown document

File extension should be `.Rmd`

```
---
title: "A Simple Regression"
author: "Yihui Xie"
output: html_document
---
We built a linear regression model.
```{r}
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```

The *slope* of the regression is **`r b[2]`**.
```

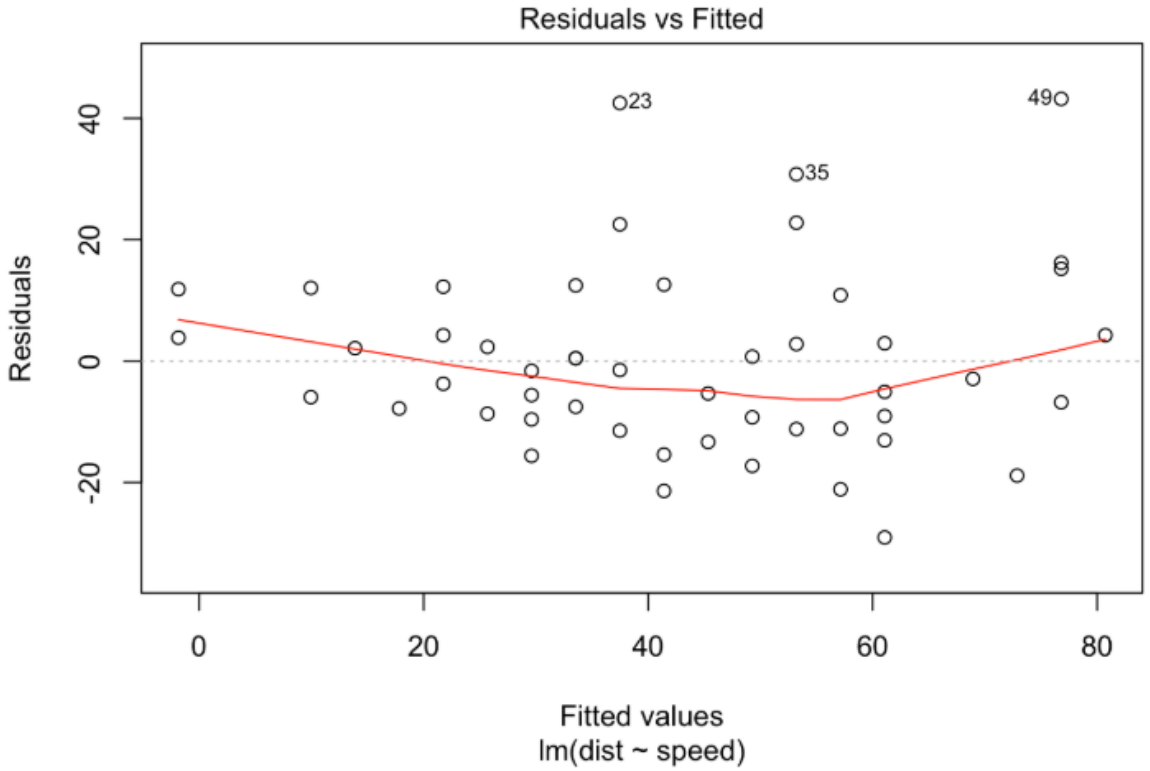
Now it's now time to start knitting 

A Simple Regression

Yihui Xie

We built a linear regression model.

```
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```



The *slope* of the regression is **3.9324088**.

An R Markdown document

File extension should be `.Rmd`

```
---
title: "A Simple Regression"
author: "Yihui Xie"
output: pdf_document
---
We built a linear regression model.
```{r}
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```

The *slope* of the regression is **`r b[2]`**.
```

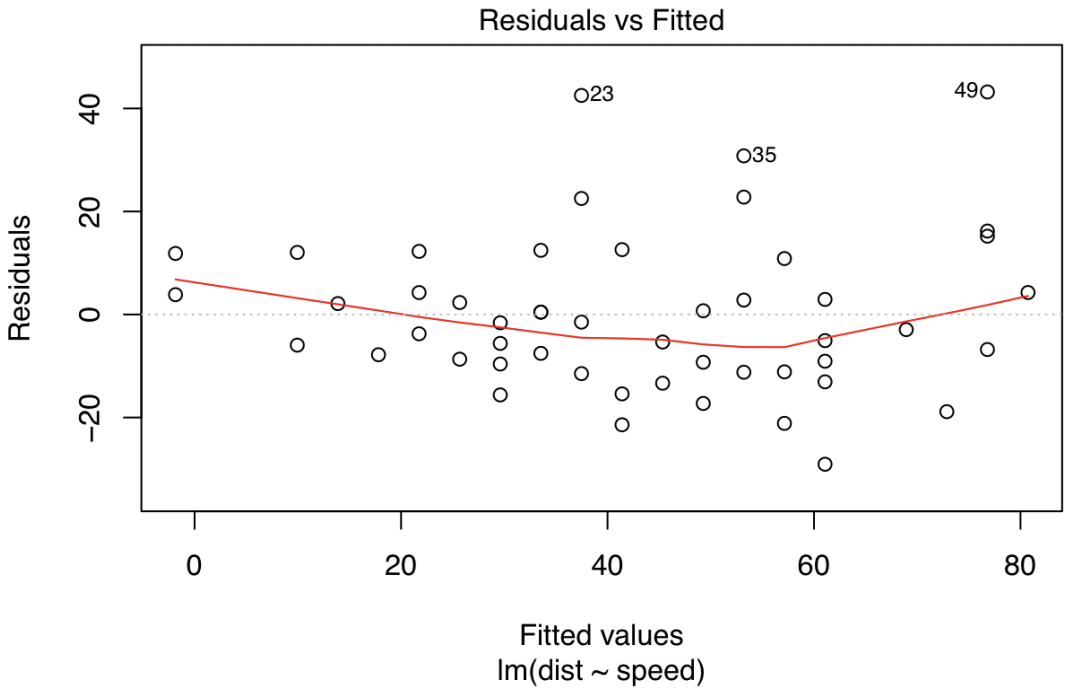
- The output of your document is now a **pdf** file.
- You will need to have LaTeX installed for this to work!
- You can get LaTeX also through tinytex R-package.

A Simple Regression

Yihui Xie

We built a linear regression model.

```
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```



The *slope* of the regression is **3.9324088**.

An R Markdown document

File extension should be `.Rmd`

```
---
title: "A Simple Regression"
author: "Yihui Xie"
output: word_document
---
We built a linear regression model.
```{r}
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```

The *slope* of the regression is **`r b[2]`**.
```

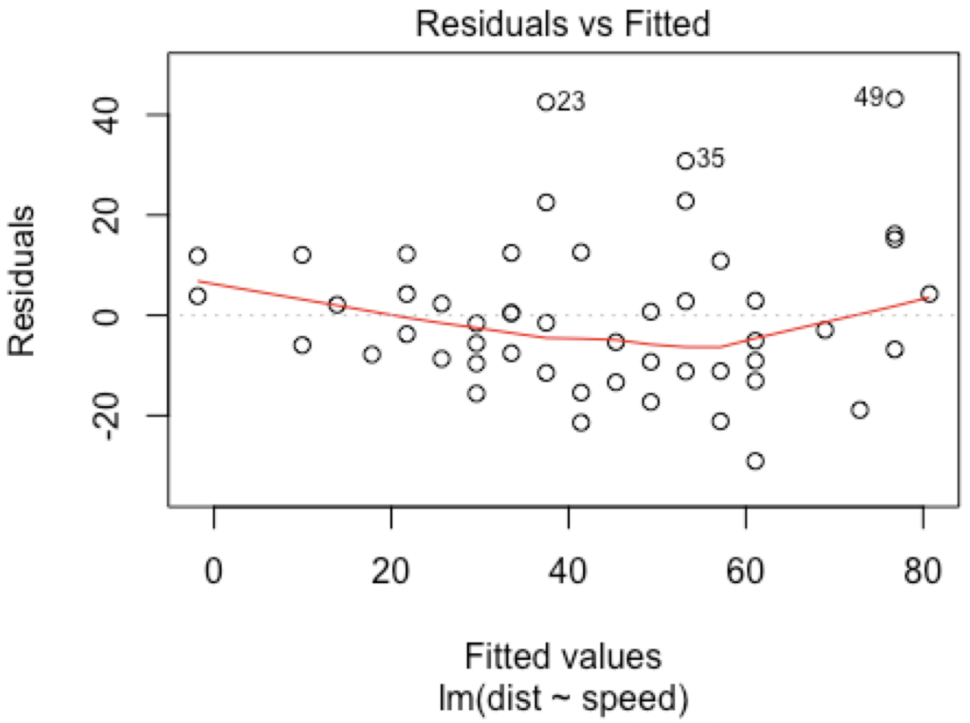
The output of your document is now a **word** file.

A Simple Regression

Yihui Xie

We built a linear regression model.

```r
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```



The *slope* of the regression is **3.9324088**.

Don't want your code chunks to appear in the report?

```{r, echo=F}
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```

Don't want your code chunks to be evaluated (but appear) in the report?

```{r, eval=F}
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```

Want your code chunks to be evaluated but hide code and output?

```{r, include=F}
fit <- lm(dist ~ speed, data = cars)
b   <- coef(fit)
plot(fit, 1)
```

You can find other code chunk options here.

R Markdown

- Fully reproducible reports -- each time you knit the analysis is ran from the beginning.

What is the difference between Markdown and R Markdown?

```
summary(cars$dist)
```

```
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.00   26.00   36.00   42.98   56.00  120.00
```

```
summary(cars$speed)
```

```
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     4.0    12.0    15.0    15.4    19.0    25.0
```

R Code Blocks

R code will be evaluated and printed

```
```{r}
summary(cars$dist)
summary(cars$speed)
```
```

Inline R Code

```
There were `r nrow(cars)` cars studied
```

Excuse me, do you have time to talk about your future?

Summary

- Carefully read the expectations.
- Be sure to schedule and plan well (important graduates attributes!)
- Get your toolkit ready - study them on your own time if you are not familiar with them.
- Remember your BIG goal - mastery of the theory and applications broadly is helpful than you knowing how to answer specific questions.
- Use R Markdown for reproducible reports.
- Learn to use R Markdown in this week's computer lab (needed for assignments).

Next lesson

- Data wrangling and visualisation in R

To do

- Download and install toolkits