The Challenges of Reproducible Research and Teaching it

Dr Chris Playford (University of Exeter)
Dr Roxanne Connelly (University of York)
Prof Vernon Gayle (University of Edinburgh)

Love Your Code, UKDS & ONS

14th February 2020
ONS London

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UNIVERSITY of York

THE UNIVERSITY OF EDINBURGH

UNIVERSITY OF EXETER

Challenges of reproducible research

- Transparency:
  - The information we present in a manuscript should enable others to fully understand and reproduce what we have done

- Challenges
  - Data sharing / data citing
  - Research Code sharing
  - Identifying the code that produced the results presented
    - Complex data management
    - Good workflow among researchers (i.e. includes all steps taken sequentially)

- Examples of how we are trying to teach better practice
  - Learning from computer science - see also Playford et al. (2016)
  - Using tools from research code-sharing. For example, see: Connelly and Gayle (2019) and supporting Jupyter Notebook

Examples of software used in research and teaching

- Literate programming / dynamic documents
  - Combining programming language with documentation language
  - Inline code

RMarkdown

- Decentralised version control
  - Platform for code sharing
  - Take a local copy of a text based file (e.g. research code)
  - Keep track of every change to a file

GitHub

1st Year
Introduction to Social Data
University of Exeter Q-step Centre

Interweaving code and outputs using RMarkdown

lab1_markdown_questions.Rmd

Knit | Insert | Run

```
Variable    | Description
----------- | -----------
year        | election year
ANES        | ANES estimated turnout rate
VEP         | voting eligible population (in thousands)
VAP         | voting age population (in thousands)
total       | total ballots cast for highest office (in thousands)
felons      | total ineligible felons (in thousands)
noncitizens | total non-citizens (in thousands)
overseas    | total eligible overseas voters (in thousands)
osvoters    | total ballots counted by overseas voters (in thousands)


## EXERCISE 1
Load the data into R and check the dimensions of the data.
Also, obtain a summary of the data. How many observations are there?
What is the range of years covered in this data set?

STEPS:

Set file path

Load your data from GitHub:

Look at your data:
```

1:1 | SOC1004/POL1008 Lab Questions | R Markdown

Variable	Description
year	election year
ANES	ANES estimated turnout rate
VEP	voting eligible population (in thousands)
VAP	voting age population (in thousands)
total	total ballots cast for highest office (in thousands)
felons	total ineligible felons (in thousands)
noncitizens	total non-citizens (in thousands)
overseas	total eligible overseas voters (in thousands)
osvoters	total ballots counted by overseas voters (in thousands)

EXERCISE 1

Load the data into R and check the dimensions of the data. Also, obtain a summary of the data. How many observations are there? What is the range of years covered in this data set?

STEPS:

Set file path

Load your data from GitHub:

Look at your data:

EXERCISE 1

Load the data into R and check the dimensions of the data. Also, obtain a summary of the data. How many observations are there? What is the range of years covered in this data set?

Load your data from GitHub:

```
# install.packages("RCurl")
# library(RCurl)
# x <- getURL("https://raw.githubusercontent.com/kosukeimai/qss/master/INTRO/turnout.csv")
# turnout <- read.csv(text = x)

# Alternatively, there are simpler ways:
# use file location:
# turnout <- read.csv("C:/Users/pe247/Documents/Q-Step teaching/POL1008/lab1/turnout.csv")
turnout <- read.csv("turnout.csv")
```

Look at your data:

```
# Check out the dimensions of your data
dim(turnout)
```

```
## [1] 14 10
```

```
# Summary of the variables in the dataset
summary(turnout)
```

```
##        X              year          VEP             VAP
##  Min.   : 1.00   Min.   :1980   Min.   :159635   Min.   :164445
##  1st Qu.: 4.25   1st Qu.:1986   1st Qu.:171192   1st Qu.:178930
##  Median : 7.50   Median :1993   Median :181140   Median :193018
##  Mean   : 7.50   Mean   :1993   Mean   :182640   Mean   :194226
##  3rd Qu.:10.75   3rd Qu.:2000   3rd Qu.:193353   3rd Qu.:209296
##  Max.   :14.00   Max.   :2008   Max.   :213314   Max.   :230872
```

2nd & 3rd Year

Data Analysis in Social Science III

University of Exeter Q-step Centre

Using GitHub classroom to share and mark assignments (Dr Alexey Bessudnov)

https://github.com/datan3-2020/datan3

Data Analysis 3: Week 6

https://github.com/datan3-2020/datan3/blob/master/class6.md

1. Read the *indresp* file from Wave 8 and keep the following variables: pidp, derived sex and age, ethnic group (h_indresp), government office region (h_gor_dv), and net personal income (h_fimnnet_dv).

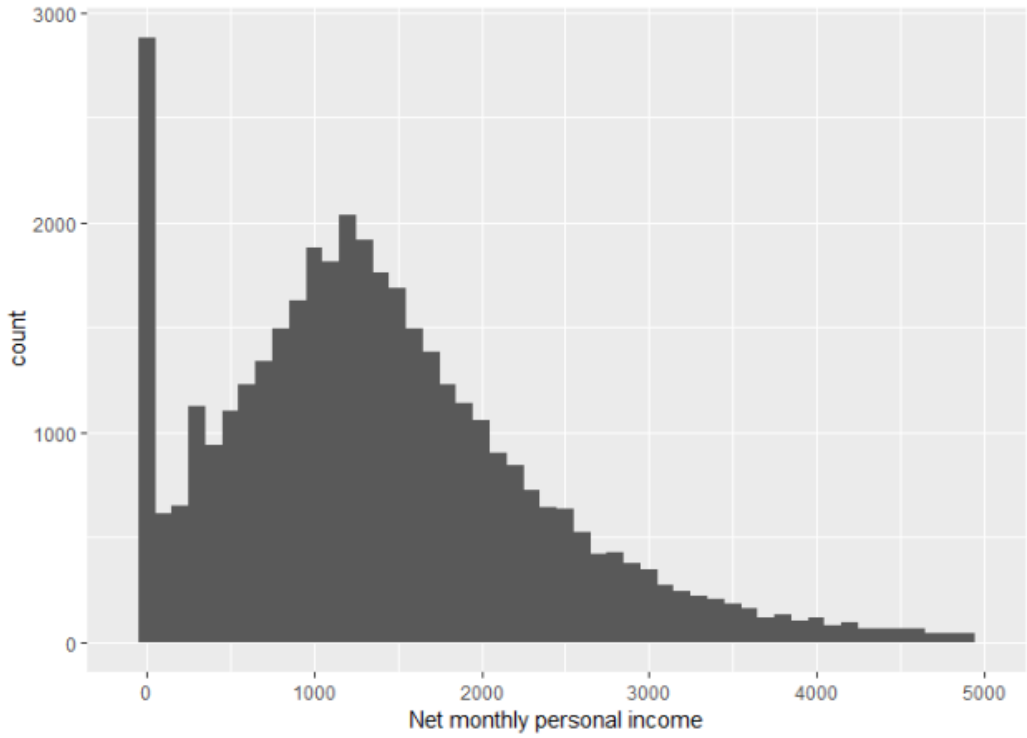
```
library(tidyverse)
Data8 <- read_tsv("data/UKDA-6614-tab/tab/ukhls_w8/h_indresp.tab")
Data8 <- Data8 %>%
  select(pidp, h_sex_dv, h_age_dv, h_gor_dv, h_fimnnet_dv)
```

For all charts use ggplot2. You may need to clean and recode variable before visualising.

We will start with univariate distributions.

1. Visualise the distribution of income with a histogram, a density plot and a box plot.

```
ggplot(Data8,
   aes(x = h_fimnnet_dv)) +
  geom_histogram(binwidth = 100) +
  xlim(-100, 5000) +
  xlab("Net monthly personal income")
```



GitHub Classroom GitHub Education

← Back to classrooms

Data analysis 3 class 2019

dataanalysis3

Assignments 7 | Students 27 | TAs and Admins 3 | Settings

Assignments

New assignment

- Final report — Individual assignment — Invite link
- Statistical assignment 1 — Individual assignment — Invite link
- Statistical assignment 2 — Individual assignment — Invite link

```
Long
```

```
@@ -84,14 +88,14 @@ Now we want to filter the data keeping only respondents from the original UKHLS
```

```{r}
Long <- Long %>%
-         filter(...) %>%
-         mutate(sex_dv = ...) %>%
-         mutate(vote6 = ...)
+         filter(memorig == 1) %>%
+         mutate(dvage = ifelse(dvage > 0, dvage, NA)) %>%
+         mutate(sex_dv =  ifelse(sex_dv == 1, "Male",
+                                 ifelse(sex_dv == 2, "Female", NA)))%>%
+         mutate(vote6 = ifelse(vote6 > 0, vote6, NA))

- Long %>%
-         ...
- Long %>%
-         ...
+ Long%>%
+   count(sex_dv, vote6)
```

abessudnov on Mar 13, 2019

20/20

Reply...

Reshape the data frame with summary statistics (20 points)

Your resulting data frame with the means is in the long format. Reshape it to the wide format. It should look like this:

sex_dv	a	b	c	d	e	f	g
female							
male							

In the cells of this table you should have mean political interest by sex and wave.

Write a short interpretation of your findings.

```
meanVote6 %>%
        gather(meanVote6, key = "variable", value = "value") %>%
        unite("variable", c("wave", "variable"), sep = "_") %>%
        spread(key = variable, value = value)
```

```
## # A tibble: 2 x 8
## # Groups:   sex_dv [2]
##   sex_dv a_meanVote6 b_meanVote6 c_meanVote6 d_meanVote6 e_meanVote6
##   <chr>        <dbl>       <dbl>       <dbl>       <dbl>       <dbl>
## 1 Female        2.84        2.82        2.87        2.89        2.87
## 2 Male          2.53        2.51        2.54        2.55        2.51
## # … with 2 more variables: f_meanVote6 <dbl>, g_meanVote6 <dbl>
```

```
# Original coding: smaller values indicate stronger interest in politics. 1 = very, 2 = fairly, 3 = not very, 4 =
# The general trend of political interest by wave remains the same for both female and male, which increased sligh
# For each wave, males tend to be more interested in politics than female.
```

Reflections

- Using dynamic documents (e.g. RMarkdown) is pretty straightforward if you are working with research code already

- Each time there is a break in the workflow, this presents opportunities for inconsistencies and error
  - Directly downloading data from an internet address is easiest
  - For data with higher levels of security, this isn't appropriate
  - Data citation and archiving really important

- Teaching social science students how to use GitHub requires patience and resources
  - You need really good teaching assistants
  - Students need good foundation in working with research code first
  - You are helping them to learn a very useful skill

Acknowledgements

- University of Exeter Q-step Centre
  - Dr Alexey Bessudnov
  - Dr Patrick English
- Dr Raluca Popp (University of Kent)

- This work was supported by the Economic and Social Research Council [ES/R004978/1]