My Final College Paper

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Wenxin Du

May 2020

Approved for the Division
(Mathematics)

Andrew Bray

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

Chapter 1

Background

1.1 Specification Curve Analysis (SCA)

1.2 Orben's Study

Chapter 2

Replication

This section discusses the attempt to replicate Orben's study along with the assessment of the use of SCA in this study. We begin by introducing the three datasets used, which can all be found through public sources under permission. We then discuss in details the attempt to replicate the study, including the obstacles to overcome during the replication process.

2.1 Data and Reprocessing

Three large-scale social datasets were used in Orben's study: Monitoring the Future (MTF) from the US [cite], Youth Risk and Behavior Survey (YRBS) from the US [cite], and Millennium Cohort Study (MCS) from the United Kingdom [cite]. The three datasets were all survey data obtained from scientific study of the same name, and encompass survey answers from adolescents aged predominately 12-18 in the time period of 2007 to 2016. The datasets provided wide measures of adolescents' psychological well-being and digital technology use. A considerable number of psychology studies in the existing literature were conducted based on the large-scale studies, which provided wide selection of approaches to modeling and analysis based on the specific dataset. In this section, we discuss the background information of the three datasets and the reprocessing of the data obtained from public sources.

2.1.1 YRBS

The Youth Risk Behavior Surveillance was first launched in 1990, and it's a biennial survey of adolescents that reflects a nationally representative sample of students attending secondary schools. Orben's study focused on the data collected during the time period of 2007 to 2015, and the same set of data was obtained through (website name)[cite]. While Orben used data in SPSS format, we were only able to access the data through Microsoft Access. The datasets were extracted and saved under excel format. It was confirmed that same number of observations were included in the obtained dataset as the data used by Orben, 37,402 girls and 37,412 boys from 2007 to 2015. It was also confirmed that all variables used in Orben's study are contained in the obtained dataset. Most of the work in the reprocessing step for YRBS focused on transforming the characteristic values of the variables used by Orben into relative numerical values.

One noticeable obstacle in this reprocessing step was, since the study is conducted anually and is still ongoing, the survey questions and indexings have been updated several times in the recent years. The majority of the variables in the datasets are named after the survey questions indexes, and the recent updates in survey questions result in differences of indices for survey questions between the current survey and surveys conducted prior to 2015. This lead to mismatches between variable names in the incorporated dataset including data from year of 2015 and prior–the one used by Orben–and the variable names in the dataset obtained for this study, including data from the year of 2017 and prior. Careful research and recoding are done to ensure the correct set of variables was used for the replication.

2.1.2 MTF

Monitoring the Future was first launched in the year of 1975, and it is an annual nationally representative survey of approximately 50,000 US adolescents in grades 8, 10 and 12. Surveys on adolescents in grade 12 were not used in the analysis since "many of the key items of interest cannot be correlated in their survey". Orben focused on the data collected during the time period of 2008 to 2016, which included 136,190 girls and 132,482 boys. The data are publicly accessible. In Orben's study, a merged dataset containing MTF data from 2008 to 2016 was used. While the MTF data for each year is publicly accessible, no access to a merged MTF dataset for the specified time period have been found. During the time period of 2008 to 2016, the survey has been updated multiple times, along with one major change in data file format after RStudio's release in the year of 2011. Due to the frequent updates in the annual surveys and changes in data files, the variable names vary greatly among the available datasets. This bring excessive difficulties to obtain the exact same dataset as used in Orben's study for replication purpose.

2.1.3 MCS

The Millennium Cohort Study follows a specific cohort of children born between September 2000 and January 2001 and collects data from both the children and the caregivers. Orben's study focused specifically on the data collected in 2015, when the participated children were aged between 13 and 15. The sample included 5926 girls and 5946 boys along with 10605 caregivers. The same dataset as used by Orben was obtained. The access to the data is open to public but require specific permission. While Orben obtained data in csv format, we were only able to obtain data in SPSS format. The same set of observations, with 5926 girls and 5946 boys borned between September 2000 and January 2001, were included in the dataset, along with the same set of variables as used in Orben's study.

Unlike working with YRBS and MTF, the variable names in the obtained dataset matches well with the variable names in the dataset used by Orben. However, instead of using numerical indices to represent survey answers, in the dataset obtained, the variable values were all in characters with the specific content refering to the specific survey answer. After careful reprocessing, all variable values were transformed into the exact numerical indices matching with the variables values as were in Orben's study. However, two variables–one related to family incomes and one related to siblings–had only NA values in the obtained dataset. The omissions might be done for confidential purpose. The two variables were used as control variables in Orben's study. As we fail to obtain the two variables, they were removed for this attempt to replication.

2.2 Replication

After obtainning the datasets we began the replication of Orben's study. The replication consists of two parts, the replication of a single SCA analysis for each dataset, and the replication of the SCA permutation test, which was used by Orben to assess the significance of the single SCA result. In the following section we discuss the procedure, obstacles and the specific resolutions to the obstacles of replicating the analysis.

2.2.1 SCA

The first part of the replication is to replicate the single SCA analysis for each dataset. All the replications in this section were done mostly by the original code provided by Orben in the public github repository. Due to the necessary reprocessings mentioned in the previous sections, slight modifications were made to the original code for the replication to be done smoothly.

It is important to understand what Orben considers as a "specification" and how a specification is identified in this study. **(Include a "definition" of specification here)** In this case, the model is set to be a linear regression, with a response variable representing "adolescent mental well-being" and an independent variable representing "technology use", with an optional set of other independent variable considered as control variables. An alternative speicification here is an alternative combination of variables to be used in the linear model. For example, one alternative specification may be using the variable "amount of time spent on watching TV in a day" as the independent variable representing "technology use", "number of times thought of suicide" as the dependent variable representing "adolescent mental well-being", and a list of selected variables as control variables, while another alternative specification choose "whether or not you own a personal computer at home" as the independent variable representing "technology use" instead. An identified specification include one identified variable for "technology use", one identified variable for "adolescent mental well-being", and making the decision of whether or not to include control variables in the model (i.e. simple linear regression or multivariate linear regression). The number of specifications determined vary among three datasets, as the number of relevant variables is different in different dataset.

Nearly 2.5 trillion alternative specifications were determined for the MCS study. Considering the computational ability, a random subset of 20,004 specifications for the SCA analysis on MCS data was used instead. A seed is not provided by Orben for the random subset, thus we failed to obtain the exact same subset of specification for this SCA analysis. We instead randomly generated our own subset of 20,004 specifications. It is noteworthy that this randomness can result in discrepancy in SCA result. Considering that the random subset has large size, we expect the degree of this discrepancy to be small. And this expectation is confirmed by replication result: while Orben obtained the median coefficient of the independent variable to be $Median(\beta) = -0.032$, our replication obtained $Median(\beta) = -0.0328$.

The problem does not exist for the studies YRBS and MTF. There were less variables available in the dataset relating to technology use and adolescent mental well-being. The number of specifications identified in the two studies are in reasonable size, therefore the exact set of specifications were used for the replications. The result matched well with Orben's result. The median coefficient of the independent variable in the YRBS study was found to be $Median(\beta) = -0.035$ in Orben's study. The result obtained in this replication, when rounded to the same digits, is also -0.035.

MTF to be discussed

2.2.2 Bootstrapping test

The next part of the replication is to replicate the inference of the single specification curves for each data. Orben chose to use a bootstrapping test on the median overall point estimate for significance of the result. We will later assess the choice of the inference test and the correctness of the inference. For now, we will focus mainly on replicating the process to conduct the bootstrapping test as Orben did and the attempt to replicate her result.

500 SCA tests were conducted in Orben's study on bootstrapped samples for each of the three datasets, and the single SCA results were shown to be significant for all three datasets. The code for the bootstrapping test and SCA are all publically available on Orben's github repository [cite]. The initial attempt of the replication was done using the original code. However, due to the large sizes of the three datasets and the great number of loops used in the R code, the replication process becomes computational expensive. A single SCA will take around 8 hours to run, and performing 500 SCA will take nearly 24 weeks. An ARC computer cluster at Oxford was used by Orben to reduce running time, however, no access to such advanced computer is available for this replication. Therefore, instead of using purely the original code, the code for single SCA and bootstrapping distribution of SCA results have been rewritten using parallel running. The running time have been significantly reduced. The dataset YRBS has the least number of observations and specifications, and after the recoding it now takes about 9 hours for a complete boostrapping test with 500 SCA's to be done using a computer with 8 cores. More time will be needed for the other two datasets, as the number of observations and specifications can be much higher in those two cases. A 96-core server is used. **Detail times should be added later**

2.3 Evaluateing Orben’s work

The replication of Orben’s work allows a better understanding of Orben’s approach and procedure. As mentioned in previous chapter, there exists a number of errors in this study in terms of the usage of SCA, including fundamental misunderstanding of the intentionals of the SCA method, inappropriate choice of specifications, and misinterpretation of the SCA results.

2.3.1

Chapter 3

Tables, Graphics, References, and Labels

3.1 Tables

In addition to the tables that can be automatically generated from a data frame in **R** that you saw in [R Markdown Basics] using the `kable` function, you can also create tables using *pandoc*. (More information is available at `http://pandoc.org/README.html#tables`.) This might be useful if you don't have values specifically stored in **R**, but you'd like to display them in table form. Below is an example. Pay careful attention to the alignment in the table and hyphens to create the rows and columns.

Table 3.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child	Inherited
Education	-0.49	Yes
Socio-Economic Status	0.28	Slight
Income	0.08	No
Family Size	0.18	Slight
Occupational Prestige	0.21	Slight

We can also create a link to the table by doing the following: Table 3.1. If you go back to [Loading and exploring data] and look at the `kable` table, we can create a reference to this max delays table too: Table **??**. The addition of the `(\#tab:inher)` option to the end of the table caption allows us to then make a reference to Table `\@ref(tab:label)`. Note that this reference could appear anywhere throughout the document after the table has appeared.

3.2 Figures

If your thesis has a lot of figures, *R Markdown* might behave better for you than that other word processor. One perk is that it will automatically number the figures accordingly in each chapter. You'll also be able to create a label for each figure, add a caption, and then reference the figure in a way similar to what we saw with tables earlier. If you label your figures, you can move the figures around and *R Markdown* will automatically adjust the numbering for you. No need for you to remember! So that you don't have to get too far into LaTeX to do this, a couple **R** functions have been created for you to assist. You'll see their use below.

In the **R** chunk below, we will load in a picture stored as `reed.jpg` in our main directory. We then give it the caption of "Reed logo", the label of "reedlogo", and specify that this is a figure. Make note of the different **R** chunk options that are given in the R Markdown file (not shown in the knitted document).

```r
include_graphics(path = "figure/reed.jpg")
```

Figure 3.1: Reed logo

Here is a reference to the Reed logo: Figure 3.1. Note the use of the `fig:` code here. By naming the **R** chunk that contains the figure, we can then reference that figure later as done in the first sentence here. We can also specify the caption for the figure via the R chunk option `fig.cap`.

Below we will investigate how to save the output of an **R** plot and label it in a way similar to that done above. Recall the `flights` dataset from Chapter **??**. (Note that we've shown a different way to reference a section or chapter here.) We will next explore a bar graph with the mean flight departure delays by airline from Portland for 2014. Note also the use of the `scale` parameter which is discussed on the next page.

```
flights %>% group_by(carrier) %>%
  summarize(mean_dep_delay = mean(dep_delay)) %>%
  ggplot(aes(x = carrier, y = mean_dep_delay)) +
  geom_bar(position = "identity", stat = "identity", fill = "red")
```

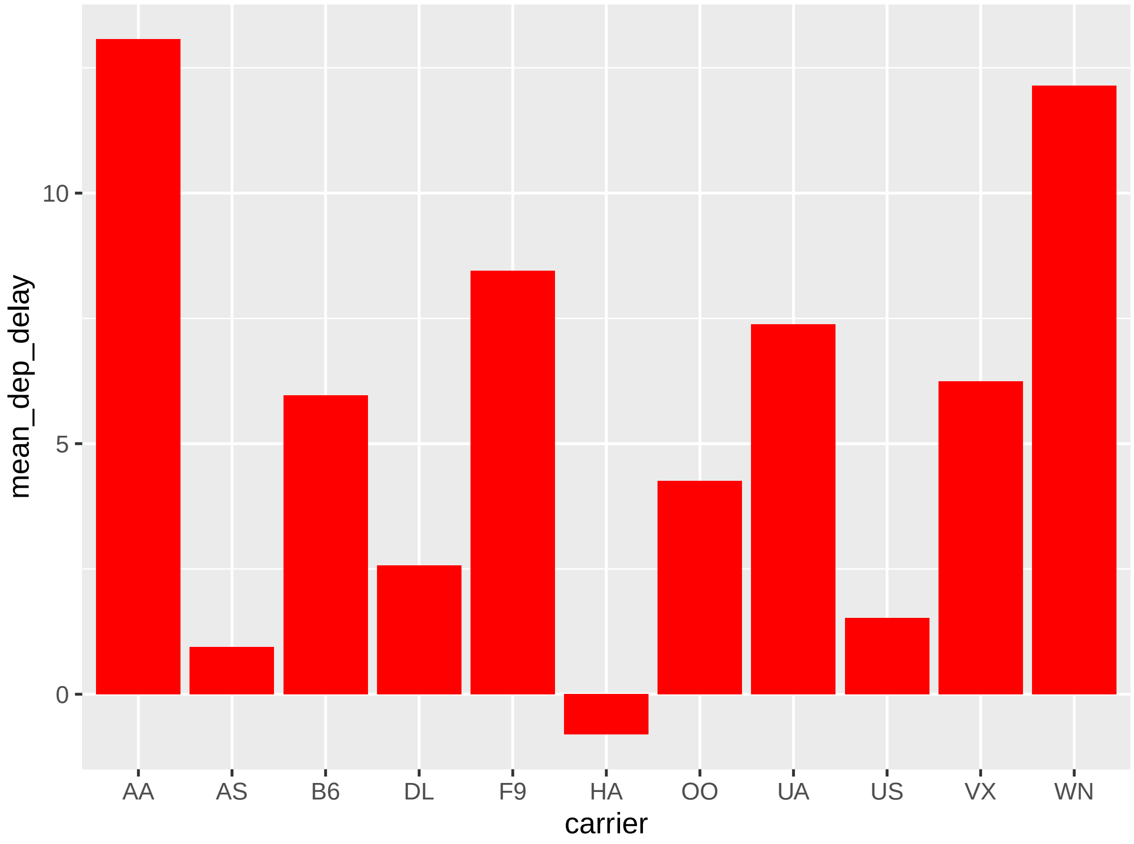

Figure 3.2: Mean Delays by Airline

Here is a reference to this image: Figure 3.2.

A table linking these carrier codes to airline names is available at `https://github.com/ismayc/pnwflights14/blob/master/data/airlines.csv`.

Next, we will explore the use of the `out.extra` chunk option, which can be used to shrink or expand an image loaded from a file by specifying `"scale= "`. Here we use the mathematical graph stored in the "subdivision.pdf" file.

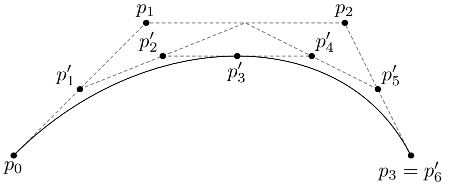

Figure 3.3: Subdiv. graph

Here is a reference to this image: Figure 3.3. Note that `echo=FALSE` is specified so that the **R** code is hidden in the document.

More Figure Stuff

Lastly, we will explore how to rotate and enlarge figures using the `out.extra` chunk option. (Currently this only works in the PDF version of the book.)

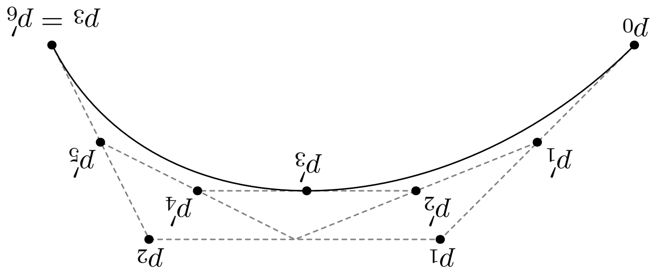

Figure 3.4: A Larger Figure, Flipped Upside Down

As another example, here is a reference: Figure 3.4.

3.3 Footnotes and Endnotes

You might want to footnote something.[1] The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be found about both on the CUS site or feel free to reach out to `data@reed.edu`.

3.4 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Reed librarians have created Zotero documentation at `http://libguides.reed.edu/`

[1] footnote text

`citation/zotero`. In addition, a tutorial is available from Middlebury College at `http://sites.middlebury.edu/zoteromiddlebury/`.

R Markdown uses *pandoc* (`http://pandoc.org/`) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: (Molina & Borkovec, 1994). This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see our CUS site (`http://web.reed.edu/cis/help/latex/index.html`)[2]. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at `http://web.reed.edu/cis/help/latex/bibtex.html`), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at `http://web.reed.edu/cis/help/latex/bibtexstyles.html`) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at `http://web.reed.edu/cis/help/latex/bibman.html`). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at `https://www.zotero.org/styles`. Make sure to download the file into the csl folder.

Tips for Bibliographies

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.
- The cite key (a citation's label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.
- Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.
- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.
- You can add a Reed Thesis citation[3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

[2] Reed College (2007)

[3] Noble (2002)

3.5 Anything else?

If you'd like to see examples of other things in this template, please contact the Data @ Reed team (email `data@reed.edu`) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(thesisdown))
  devtools::install_github("ismayc/thesisdown")
library(thesisdown)
```

In Chapter 3:

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(dplyr))
    install.packages("dplyr", repos = "http://cran.rstudio.com")
if(!require(ggplot2))
    install.packages("ggplot2", repos = "http://cran.rstudio.com")
if(!require(ggplot2))
    install.packages("bookdown", repos = "http://cran.rstudio.com")
if(!require(thesisdown)){
  library(devtools)
  devtools::install_github("ismayc/thesisdown")
  }
```

```r
library(thesisdown)
flights <- read.csv("data/flights.csv")
```

Appendix B

The Second Appendix, for Fun

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl*. Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime*. Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.

Molina, S. T., & Borkovec, T. D. (1994). The Penn State worry questionnaire: Psychometric properties and associated characteristics. In G. C. L. Davey & F. Tallis (Eds.), *Worrying: Perspectives on theory, assessment and treatment* (pp. 265–283). New York: Wiley.

Noble, S. G. (2002). *Turning images into simple line-art* (Undergraduate thesis). Reed College.

Reed College. (2007, March). LaTeX your document. Retrieved from `http://web.reed.edu/cis/help/LaTeX/index.html`