Course Final Project

This project serves as an opportunity to apply the techniques presented in the course videos to attempt to reproduce the findings of a recently published journal article:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.

In 1995, MIDUS survey data were collected from a total of 7,108 participants. The baseline sample was comprised of individuals from four subsamples: (1) a national RDD (random digit dialing) sample (n=3,487); (2) oversamples from five metropolitan areas in the U.S. (n=757); (3) siblings of individuals from the RDD sample (n=950); and (4) a national RDD sample of twin pairs (n=1,914). All eligible participants were non-institutionalized, English-speaking adults in the contiguous United States, aged 25 to 74. All respondents were invited to participate in a phone interview of approximately 30 minutes in length and complete 2 self-administered questionnaires (SAQs), each of approximately 45 pages in length. In addition, the twin subsample was administered a short screener to assess zygosity and other twin-specific information. With funding provided by the National Institute on Aging, a longitudinal follow-up of MIDUS I began in 2004. Every attempt was made to contact all original respondents and invite them to participate in a second wave of data collection. Of the 7,108 participants in MIDUS I, 4,963 were successfully contacted to participate in another phone interview of about 30 minutes in length. MIDUS II also included two self-administered questionnaires (SAQs), each of about 55 pages in length, which were mailed to participants. The overall response rate for the SAQs was 81%. Over 1,000 journal articles have been written using MIDUS I and II data since 1995.

You will attempt to reproduce the findings of [1] and critique the reproducibility of the article. This particular article focuses only on MIDUS II data, including biomarker data, and investigates the relationship between optimism and lipids. You can download the MIDUS II data and supporting codebook and other documents here. You can download the data in multiple formats. You can download the biomarker data here.

The project must be submitted in the form of a Jupyter notebook or RMarkdown file, and include an introduction, methods, results, and discussion/critique section with commented code interspersed. Visuals and schematics should be included if appropriate. All project documents must also be uploaded to the course GitHub repository in a folder. The folder must also include a README file describing the contents of the folder and how to reproduce all results. You should keep in mind the file and folder structure covered in the videos and make the process as automated as possible.

1. **Is the data publicly available?**

The data is publicly available and available in two formats: one hosted on the Institute for Social Research of University of Michigan and the second hosted on the National Archive of Computerized Data on Aging. This study is maintained and distributed by the National Archive of Computerized Data on Aging (NACDA), the aging program within ICPSR. NACDA is sponsored by the National Institute on Aging (NIA) at the National Institutes of Heath (NIH).

2. **Is the data easy/intuitive to access?**

The data is intuitive to access. The data is publicly available and wholly accessible. An overview of the data is described in 5 subheadings that provide a summary of the project and its implications, as well as related publications and accessible formats for data export. Each subheading is further explained in subsequent summaries for specific subjects.

3. **Is there a codebook and/or instructions about how the data and documentation is organized?**

The repository hosted on the ISRUM is compiled as the ICPSR_04652. The repository hosted on the NACDA is compiled as the ICPSR-29282. Each repository has the same file structure. The Manifest text file lists the project title, affiliated scientists, and study level documentation that thoroughly records 5 main descriptive variables: record length, record count, variable count, data updated and MD5 checksum. Files are organized by study-level and aggregated data. The repository also hosts a README first memo that thoroughly details data and documentation for the entire MIDUS longitudinal project, nomenclature of files, standalone instrument files and their associated description and use case.

4. **Are the file names intuitive?**

The documentation files described below are available as PDF files through the Colectica Portal and at ICPSR. The Portal supports the naming system below, but unfortunately, the file management system in place at ICPSR renames the files into the following format: Documentation.pdf (shortfilename)

The shortfilename is based on the file names of the documents we submit (see below), thus, the name of this readme file at ICPSR is something like “Documentation.pdf (readme)”. To find documents of interest on the ICPSR site it is recommended that you review the following descriptions and then look for key words from these file names in the parenthetical

shortfilenames. After downloading the files it may be helpful to rename them according to the conventions below for future reference.

Files names are intuitive at the study level and aggregated data. Study level file names are named as follows: Study acronym_Study subset unique ID_File type_Year_Month_Day.file extension.

5. **Are the variable names intuitive?**

Variable names are not intuitive. For example B1SA54C refers the how many times one saw a psychologist or professional counselor in the past 12 months about emotional or mental health. All variable names are uniquely identified but lack intuitive identification. Identification requires a one-to-one relationship table that links one unique identifier name to one variable.

6. **Describe the software. Is the software used for analysis publicly available?**

The software is known as SAS Analytics. SAS is a Business Intelligence tool that facilitates analyses, reporting, data mining, and predictive modeling with the help of powerful visualizations and interactive dashboards. SAS (Statistical Analysis System) introduced the SAS Business Intelligence and Analytics Solution for helping large enterprises explore their large datasets in a visually appealing format. SAS analytics is a data analytics tool that is used increasingly in Data Science, Machine Learning, and Business Intelligence applications. Not only it equips organizations with all necessary tools to monitor the key BI metrics but also produces powerful analytics and comprehensive reports for their decision makers to take well-informed decisions. The software is not publicly available and runs on a subscription based model.

7. **If the software is available, is it well commented?**

The software is not publicly available without a subscription.

8. **Is there a toy example provided?**

No.

9. **Are you able to reproduce the figures, tables and results presented in the paper?**

No.

10. **Was there anything you think should have been made clearer, or explained in a different way?**

A clear limitation of the present investigation is the cross-sectional data. Whether optimism leads to healthier lipid profiles or whether healthier lipid profiles (and better health in general) lead to optimism cannot be determined, although optimism was often measured at least several months prior to lipid profiles. Notably, optimism did not vary substantially according to whether individuals were taking lipid medications, reducing somewhat the concern that lipid levels determine optimism. Although it is suspected that optimism does in fact influence lipid levels as an upstream determinant, it is also possible that the association is bidirectional. Moreover, an unmeasured third variable could also determine both optimism and lipid profiles. Future prospective and experimental research is needed to more clearly establish the direction of effects, investigate whether associations differ depending on health status, and examine whether controlling other measures of ill-being alters findings.

11. **Did you find any faults in the methods used in this paper? Would you have used more or different methods?**

Strengths of the research include a well-validated measure of optimism and objectively-measured lipids, which limits concerns regarding self-report bias. Additional strengths include the ability to consider potential confounding and pathway variables. Taken together, this research suggests that an optimistic outlook is related to a healthier lipid profile. Thus, considering optimism in the context of lipids may suggest new strategies for prevention and intervention to improve cardiovascular health. Given the robust methodology of the paper, I would have used the same statistical, analytical and computational approach.

However, as part of any reproducible workflow, the associated statistical analysis program should be included if it is publicly available. In this case of paid, inaccessible software, an sample file of processed data at each stage of the workflow with intuitive, documented file names and clear descriptions.