group_midus

Dependencies

This notebook can be reproduced by installing the following R packages: - knitr - janitor - dplyr - tidyverse - ggplot2 - DT - tableone - table1 - gt

Reproducibility group project, BST270 2024

Introduction

In this Rmarkdown file we will attempt to reproduce the figures, tables and analyses presented in the paper *Relation between Optimism and Lipids in Midlife*.

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. http://doi.org/10.1016/j.amjcard.2013.01.292

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.

Here we 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. The MIDUS II data and supporting codebook and other documents can be downloaded here. The data can be downloaded in multiple formats. The biomarker data can be downloaded here.

Data Dictionary

This manuscript uses several variables from multiple data files. Some of these variables don't have intuitive names and need to be manually looked up either online or in the codebooks provided in the data downloads. We generated a data dictionary to our understanding of the naming conventions.

Load packages

```
library(dplyr)
library(tidyverse)
library(ggplot2)
library(DT)
```

We are trying to keep all functions well documented. This command allows us to have package-like documentation for all of the functions.

```
if (!require('devtools')) install.packages('devtools')
```

```
## Loading required package: devtools
```

```
## Loading required package: usethis
```

```
devtools::document()
```

```
## i Updating bst270Midus2024 documentation
```

```
## i Loading bst270Midus2024
```

```
## [1] "hello"
```

```
## Warning: -- Conflicts ---------------------------------------- bst270Midus2024 conflicts
## --
## x `gen_table2` masks `bst270Midus2024::gen_table2()`.
## i Did you accidentally source a file rather than using `load_all()`?
##   Run `rm(list = c("gen_table2"))` to remove the conflicts.
```

Read data

First we load the data. 29282-0001-Data contains the analysis-associated data, while 04652-0001-Data contains the midus clinical data.

```
getwd()
```

```
## [1] "/Users/violafanfani/Documents/uni-harvard/courses/bst270/bst270-winter2024/group_project/bst270
```

```
load('data/29282-0001-Data.rda')
load('data/04652-0001-Data.rda')
```

We have to merge the two tables based on the MIDUS II ID number

```
data = inner_join(da04652.0001, da29282.0001, by = c("M2ID", "M2FAMNUM"),suffix = c('','.2'))
print(dim(data))
```

```
## [1] 1054 5415
```

Data has 1054 rows at the beginning after merging the two tables. Now we are going to try and reproduce the preprocessing steps such that we can obtain the 990 individuals they used for the paper analysis.

Wrangle data

Step 0. Filter optimism variables

Optimism is assessed using the 6-item Life-Orientation test. In the codebook we have found that the B1SORIEN column contains the test value, already processed as explained in the paper while columns ('B1SORIEN','B1SE10A', 'B1SE10B', 'B1SE10C', 'B1SE10D', 'B1SE10E','B1SE10F') contain the single item scores.

Here we are filtering the rows to remove the individuals who do not have an optimism score.

```
data_after_fo = data %>% filter_optimism()
nrow(data_after_fo)
```

```
## [1] 1050
```

We are left with 1050 samples. However, by visually inspecting the table we can see that for some patients, although we have a B1SORIEN value, some of the items are reported as NAs. This seems weird, we do not understand how they computed the final score, hence we proceed to remove also these patients.

```
optimism_columns <- c('B1SORIEN','B1SE10A', 'B1SE10B', 'B1SE10C', 'B1SE10D', 'B1SE10E','B1SE10F')
data_after_fo2 <- data_after_fo %>%
    drop_na(any_of(optimism_columns))
nrow(data_after_fo2)
```

```
## [1] 1041
```

We are left with 1041 samples.

Secondly, we clean the columns relative to lipids (Total cholesterol, HDL, LDL, triglicerydes).

Step 1 Filter lipid measurements

We will focus on 24 columns of our interests and drop rows with missing values in "B4BCHOL", "B4BTRIGL", "B4BHDL" and "B4BLDL" columns.

```
lipid_columns = c("B4BCHOL", "B4BTRIGL", "B4BHDL", "B4BLDL")
data_after_fl = data_after_fo2 %>% filter_lipid()
nrow(data_after_fl)
```

```
## [1] 1030
```

After filtering lipid measurements, we have 1030 rows left

We also have to clean and filter the pathway variables which are those relative to lifestyle and concern drinking, smoking and diet.

Step 2 Filter pathway variables

```
data_after_fp = filter_pathway(data_after_fl)
nrow(data_after_fp)
```

```
## [1] 1020
```

```
pathway_columns = c("B4PBMI", "alcohol_consumption", "smoking_status", "reg_exercise","score_sum")
```

We are left with 1020 samples. This filtering was tricky because we found that many choices are not adequately explained in the paper: what columns are used for the drinking habits?

Step 3 Filter potential confounders

Finally we filter the potential confounders, such as age, sex, income..

```
confounder_columns = c('B1PB1','B1STINC1','B4ZB1SLG','B1SCHROX','B4H26','B4H33','B4H25','B4PBMI','B1SNE

data_after_fc = filter_confounders(data_after_fp, confounder_columns)
nrow(data_after_fc)
```

```
## [1] 994
```

Here we are left with 994 individuals, 4 more than those reported in the paper.

Although we are not able to completely reproduce the data wrangling steps, we are now going to reproduce some of the figures and tables and assess whether the general trends reported in the paper are robust to this slight change in the population make-up.

Let's keep only the columns of interest

```
all_columns = c(optimism_columns, lipid_columns, pathway_columns, confounder_columns)
#View(data_after_fc[,all_columns])
```

Figure 1

First, we attempt to reproduce Figure 1. Figure 1 shows the frequency distribution of 990 optimism scores (mean +- SD: 23.95 +- 4.69), with black representing the lowest tertile of optimism (6 to 22), gray, middle tertile of optimism (23 to 26), and white, highest tertile of optimism (27 to 30)

In our data the average and standard deviation might be slightly different

```
mean(data_after_fc$B1SORIEN)
```

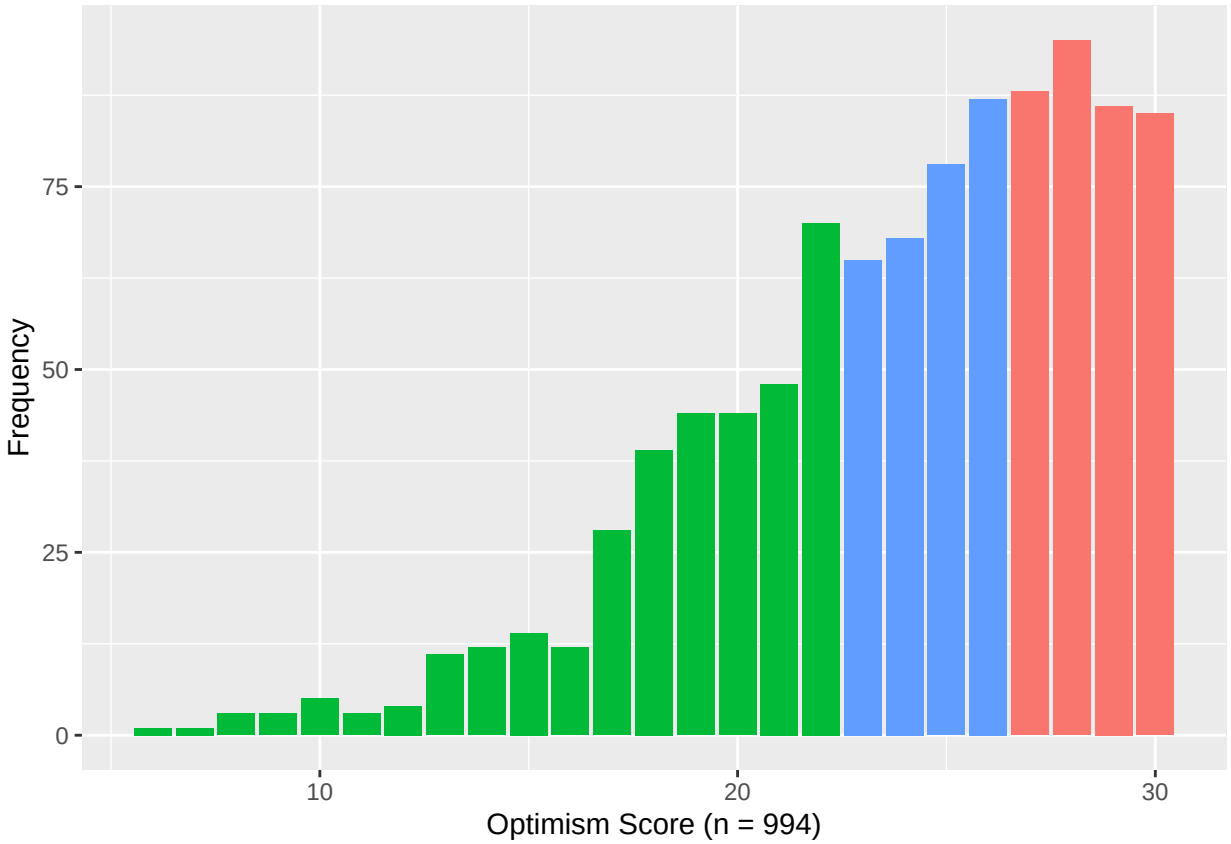
```
## [1] 23.91247
```

```
sd(data_after_fc$B1SORIEN)
```

```
## [1] 4.690779
```

Indeed we have a slight deviation in mean, that is now 23.91 vs the 23.05 reported in the table. Standard deviation is instead consistent with what was reported.

```
gen_fig1(data_after_fc)
```



The histogram of the optimism scores seems to be slightly different, for instance there seem to be more samples with score 22.

Table 1

We then proceed to reproduce table 1. We are gonna split it in different chunks, based on the lipid/confounder/pathway groups.

```
gen_table1_lipid(data_after_fc)
```

Table 1: Distribution of participant lipid measurements according to optimism level

	Low	Moderate	High
Total cholesterol	188.18+-38.4	185.36+-38.41	187.17+-39.76
HDL cholesterol	52.68+-17.44	53.88+-16.69	57.21+-18.16
LDL cholesterol	108.91+-34.72	105.33+-34.9	105.47+-35.07
Triglycerides	135.04+-83.47	132.6+-84.16	122.99+-68.46

We also reproduce the table for the pathway variables

```
gen_table1_pathway(data_after_fc)
```

```
##                                  Stratified by Optimism
##                                   Low           Moderate      High
##   n                                 342           298           354
##   BMI (mean (SD))                 29.70 (6.60)  28.91 (5.53)  28.91 (5.83)
##   Alcohol Consumption (mean (SD))  1.35 (1.40)   1.11 (1.23)   1.15 (1.39)
##   Smoking status (%)
##      current smoker                61 (17.8)     26 ( 8.7)     25 ( 7.1)
##      never smoker                 178 (52.0)    163 (54.7)    215 (60.7)
##      past smoker                  103 (30.1)    109 (36.6)    114 (32.2)
##   Regular Exercise = Yes (%)      254 (74.3)    250 (83.9)    285 (80.5)
##   Prudent Diet (mean (SD))       3.86 (1.44)   4.22 (1.33)   4.46 (1.37)
```

Finally we reproduce the table for the confounders

```
table1_function(data_after_fc)
```

```
## Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

	Total	Low	Moderate	High
	(N=994)	(N=342)	(N=298)	(N=354
Education				
Less than a high school degree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
High school degree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Some college	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
College degree or more	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Missing	994 (100%)	342 (100%)	298 (100%)	354 (100
Household_Income				
Mean (SD)	77000 (± 60000)	68000 (± 55000)	76000 (± 60000)	86000 (&
Chronic_Conditions				
No	227 (23 %)	64 (19 %)	61 (20 %)	102 (29
Yes	767 (77 %)	278 (81 %)	237 (80 %)	252 (71
Smoking				
No	556 (56 %)	178 (52 %)	163 (55 %)	215 (61
Yes	438 (44 %)	164 (48 %)	135 (45 %)	139 (39
Alcohol				
Mean (SD)	0.68 (± 0.47)	0.70 (± 0.46)	0.66 (± 0.47)	0.67 (&p
Regular_Exercise				
No	205 (21 %)	88 (26 %)	48 (16 %)	69 (19 %
Yes	789 (79 %)	254 (74 %)	250 (84 %)	285 (81
BMI				
Mean (SD)	29 (± 6.0)	30 (± 6.6)	29 (± 5.5)	29 (&plu

```
gen_table2(data_after_fc)
```

Characteristic r p

```
get_table2_confounders(data_after_fc)
```

```
##              Characteristic           r
## B1PB1             Education  0.12370051
## B1STINC1   Household Income  0.14287292
## B4ZB1SLG               LAG  0.00575478
## B1SCHROX Chronic Conditions -0.12107926
## B4H26              Smoking -0.06814463
## B4H33              Alcohol -0.02708276
## B4H25      Regular Exercise  0.06155039
## B4PBMI                 BMI -0.06944342
## B1SNEGAF    Negative affect -0.49820848
```

Critique of this study reproducibility

1. Is the data publicly available? Yes
2. Is the data easy/intuitive to access? Yes
3. Is there a codebook and/or instructions about how the data and documentation is organized? Yes, but it was not always clear
4. Are the file names intuitive? No
5. Are the variable names intuitive? No
6. Is the software used for analysis publicly available? No, they use SAS v9.2 that is not available, and

they do not give details of what they did on SAS

7. If the software is available, is it well commented? No
8. Is there a toy example provided? No
9. Are you able to reproduce the figures, tables and results presented in the paper? The variables seem to be all there, but we had a hard time matching them 1-1 with what was done in the paper. We got to 994 samples, instead of 990 and for some of the variables it looks like we have different distributions of the values.
10. Was there anything you think should have been made clearer, or explained in a different way?
11. Did you find any faults in the methods used in this paper? Would you have used more or different methods? The alcohol consumption, reported as drinks/week, but based on the reproduction steps it looks like they use drinks/day of drinking. The blood pressure medication variables were not clear We did not understand if the lipid levels had to be corrected, or were already corrected in the dataset