group_midus

Dependencies

This notebook can be reproduced by installing the following R packages: - knitr - dplyr ...

And by using the functions in the following files

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
```

Read data

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

```
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 ...

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

```
# filter optimism variables
source("filter_optimism.R")
data_after_optimism_filtering <- Filter_optimism(data)
```

We are left with x samples

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

Step 1 Filter lipid measurements

```
# filter and clean lipid measurement columns
source("filter_lipids.R")
data_after_lipid_filtering <- Filter_lipids(data)
print(dim(data_after_lipid_filtering))
```

```
## [1] 1043    5
```

After filtering lipid measurements, we have 1043 rows left

Step 2 Filter pathway variables

```r
# filter pathway varibales
source("filter_pathway.R")
data_after_pathway_filtering <- filter_pathway(data)
print(dim(data_after_pathway_filtering))
```

```
## [1] 1048    6
```

We are left with x samples

Step 3 Filter potential confounders

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

```r
# filter confounders
source("filter_confounders.R")
data_after_confounder_filtering = Filter_confounders(data)
print(dim(data_after_confounder_filtering))
```

```
## [1] 999  10
```

Here we are left with x individuals

Let's keep only the columns of interest

```r
full_data <- inner_join(data_after_optimism_filtering,
                        data_after_lipid_filtering, by = "M2ID") %>%
    inner_join(data_after_confounder_filtering, by = "M2ID") %>%
    inner_join(data_after_pathway_filtering, by = 'M2ID')

print(dim(full_data))
```

```
## [1] 977  20
```

```r
# 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)

```r
#generate figure 1
```

Table 1

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

```r
# generate table 1
```

Table 2

We then proceed to reproduce table 2. They don't specify how the correlations and p-values are calculated, so we'll assume Spearman correlation and HC3 standard errors

```r
fig2_cols <- c("Optimism", "Age", "Gender", "Race", "Education", "Income",
               "Interval between assessments", "Chronic conditions",
```

```
               "Blood pressure medication", "Body mass index", "Smoking status",
               "Alcohol consumption", "Prudent diet", "Regular exercise",
               "Negative affect")
cor_data <- full_data %>%
    mutate(Optimism = B1SORIEN,
           Age = as.numeric(age),
           Gender = case_when(sex == '(1) Male' ~ 0,
                              sex == '(2) Female' ~ 1),
           Race = case_when(race == 'White' ~ 0,
                            race == 'Nonwhite' ~ 1),
           Education = education_categorical,
           Income = household_income,
           `Interval between assessments` = visit_interval,
           `Chronic conditions` = case_when(chronic_condition == '(0) No' ~ 0,
                                            chronic_condition == '(1) Yes' ~ 1),
           `Blood pressure medication` = case_when(blood_pressure_med == '(2) No' | blood_pressure_med =
                                                   blood_pressure_med == '1' ~ 1),
           `Body mass index` = BMI,
           `Smoking status` = case_when(smoking_status == '(1) current smoker' ~ 1,
                                        smoking_status == '(2) past smoker' ~ 2,
                                        smoking_status == '(3) never smoker' ~ 3),
           `Alcohol consumption` = drinks_per_day,
           `Prudent diet` = prudent_diet_score,
           `Regular exercise` = case_when(regular_exercise == '(2) No' ~ 0,
                                          regular_exercise == '(1) Yes' ~ 1),
           `Negative affect` = negative_affect) %>%
    select(all_of(fig2_cols))

results_df <- data.frame(matrix(ncol = 3, nrow = length(fig2_cols) - 1))
colnames(results_df) <- c("Characteristic", "r", "p")
for (i in seq(fig2_cols[-1])) {
    colname <- fig2_cols[-1][i]
    cor_val <- c(cor(cor_data[colname], cor_data$Optimism, method = 'pearson'))
    lm_model <- lm(cor_data$Optimism ~ unlist(cor_data[colname]))
    pval <- coef(summary(lm_model))[2, 'Pr(>|t|)']
    results_df$Characteristic[i] <- colname
    results_df$r[i] <- cor_val
    results_df$p[i] <- pval
}

results_df <- results_df %>%
    mutate(r = round(r, 2),
           p = case_when(p < 0.0001 ~ "<0.0001",
                         p < 0.001 ~ "<0.001",
                         TRUE ~ as.character(round(p, 2))))

knitr::kable(results_df)
```

Characteristic	r	p
Age	0.20	<0.0001
Gender	0.01	0.65
Race	-0.05	0.11
Education	0.15	<0.0001

Characteristic	r	p
Income	0.11	<0.001
Interval between assessments	0.00	0.95
Chronic conditions	-0.13	<0.0001
Blood pressure medication	0.05	0.12
Body mass index	-0.07	0.02
Smoking status	0.14	<0.0001
Alcohol consumption	-0.03	0.3
Prudent diet	0.21	<0.0001
Regular exercise	0.07	0.04
Negative affect	-0.45	<0.0001

Table 5

```
if (nrow(cor_data) != nrow(full_data)) {
    stop("nrow(cor_data) != nrow(full_data)")
}

tab5_data <- cor_data %>%
    mutate(total_chol_binary = full_data$B4BCHOL >= 240,
           high_ldl_binary = cor_data$Gender == 0 & full_data$B4BHDL < 40 |
               cor_data$Gender == 1 & full_data$B4BHDL < 50,
           low_ldl_binary = full_data$B4BLDL >= 160,
           trigl_binary = full_data$B4BTRIGL >= 200,
           optimism_sd = scale(Optimism))

lipid_cols <- c("total_chol_binary", "high_ldl_binary", "low_ldl_binary", "trigl_binary")
tab5_final_results <- data.frame(matrix(ncol = 6, nrow = 0))
colnames(tab5_final_results) <- c("Lipid", "Model", "OR", "LB", "UB", "pval")
for (lipid_col in lipid_cols) {
    tmp_tab5_data <- tab5_data
    colnames(tmp_tab5_data)[colnames(tmp_tab5_data) == lipid_col] <- 'outcome'
    for (model in c("model_1", "model_2")) {
        if (model == 'model_1') {
            fit_model <- glm(outcome ~ Age + Gender + Race + Education + Income +
                   `Interval between assessments` + optimism_sd,
                    tmp_tab5_data, family = 'binomial')
        } else {
            fit_model <- glm(outcome ~ Age + Gender + Race + Education + Income +
                   `Interval between assessments` + `Chronic conditions` +
                     `Blood pressure medication` + optimism_sd,
                    tmp_tab5_data, family = 'binomial')
        }
        point_estimate <- exp(coef(fit_model))['optimism_sd']
        conf_int <- exp(confint(fit_model))['optimism_sd',]
        tab5_final_results <- rbind(tab5_final_results,
            data.frame(Lipid = lipid_col,
                       Model = model,
                       OR = point_estimate,
                       LB = conf_int[1],
                       UB = conf_int[2],
                       pval = coef(summary(fit_model))['optimism_sd','Pr(>|z|)']))
```

```
    }
}
```

```
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
```

```
tab5_final_results <- tab5_final_results %>%
    mutate(presented_name = paste0(round(OR, 2),
                                   case_when(pval < 0.05 ~ '‖',
                                             pval < 0.10 ~ '§',
                                             TRUE ~ ''),
                                   " (", round(LB, 2), "-", round(UB, 2), ")"))

tab5_out <- reshape2::dcast(tab5_final_results, Lipid ~ Model, value.var = 'presented_name')
tab5_out <- tab5_out %>%
    mutate(Lipids = case_when(Lipid == 'high_ldl_binary' ~ 'High-density lipoprotein cholesterol',
                              Lipid == 'low_ldl_binary' ~ 'Low-density lipoprotein cholesterol',
                              Lipid == 'total_chol_binary' ~ 'Total cholesterol',
                              Lipid == 'trigl_binary' ~ 'Triglycerides')) %>%
    rename(`Model 1` = model_1,
           `Model 2` = model_2) %>%
    arrange(factor(Lipid, levels = c('Total cholesterol',
                                     'High-density lipoprotein cholesterol',
                                     'Low-density lipoprotein cholesterol',
                                     'Triglycerides'))) %>%
    select(Lipids, `Model 1`, `Model 2`)

knitr::kable(tab5_out)
```

Lipids	Model 1	Model 2
High-density lipoprotein cholesterol	0.84‖ (0.73-0.97)	0.85‖ (0.73-0.98)
Low-density lipoprotein cholesterol	1.04 (0.81-1.34)	1.04 (0.81-1.34)
Total cholesterol	0.9 (0.73-1.12)	0.91 (0.73-1.13)
Triglycerides	0.87 (0.73-1.05)	0.85§ (0.7-1.03)