Biost 578: Problem Set 2

Department of Biostatistics @ University of Washington

Alejandro Hernandez

Due Friday 3, May 2024

This practice is organized around the causal question "Does being physically active cause you to live longer?" We will practice the methods we have learned (optimal multivariate matching, outcome regression, IPW, and AIPW, balancing estimators) using data from NHANES I Epidemiologic Follow-up Study. We will also perform sensitivity analyses to gauge the robustness of the evidence to possible unmeasured confounding.

The dataset nhanesi class `dataset.csv` is posted on the course website. More detail of the data can be found in the paper by Davis et al. (1994), "Health behaviors and survival among middle aged and older men and women in the NHANES I Epidemiologic Follow-Up Study."

The NHANES I sample was interviewed in 1971 and followed for survival until 1992. Physical activity was measured in two variables: self-reported nonrecreational activity and self-reported recreational activity. We consider the treatment to be adults who reported themselves to be "quite inactive", both at work and at leisure, and we will compare them to controls who were quite active ("very active" in physical activity outside of recreation and "much" or "moderate" recreational activity). The treatment variable is physically.inactive. Following Davis et al. (1994), we excluded people who were quite ill at the time of the NHANES I survey. We included people aged between 45 and 74 at baseline, and excluded people who, prior to NHANES I, had heart failure, a heart attack, stroke, diabetes, polio or paralysis, a malignant tumor, or a fracture of the hip or spine.

The measured confounders are the following:

- sex
- smoking status (current smoker, former smoker or never smoker)
- income.poverty.ratio: ratio of household income to poverty line for the household size, where this variable is top coded (right censored) at 9.98 (i.e., if is greater than 9.98, it is coded as 9.98).
- age at time of interview
- race (white vs. non-white)
- education (¡=8 years, 9-11 years, high school graduate but no college, some college, college graduate)
- working.during.last.three months – employed or not during the previous three months
- marital status
- alcohol consumption (never, ¡1 time per month, 1-4 times per month, 2+ times per week, just about everyday
- dietary adequacy (number of five nutrients – protein, calcium, iron, Vitamin A and Vitamin C – that were consumed at more than two thirds of the recommended dietary allowance)

The outcome of interest is years.lived.since.1971.up.to.1992, the number of years the person was alive between the interview in 1971 up until 1992 (the maximum value is 21 since followup ended in 1992).

1

`income.poverty.ratio` and `dietary.adequacy` have missing values (indicated by NA). Create indicator variables for whether `income.poverty.ratio` and `dietary.adequacy` have missing values and fill in the missing values with the mean of the observed values. [Note that education has a few missing values but Missing is already coded as a category for education].

2

3

4

5

6

8

9

10

Bonus

Code Appendix

```
# clear environment
rm(list=ls())

# load relevant packages
library(MASS)        # negative binomial distribution
library(tidyverse)   # data manipulation
# devtools::install_github("mbannick/RobinCar")
library(RobinCar)    # causal models with covariate adjustment
library(rigr)        # regression
library(knitr)       # table formatting

# setup options
knitr::opts_chunk$set(echo = F, warning = F)
options(knitr.kable.NA = '-')
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "llm_appendix", "allcode")]
set.seed(0927)
```

End of document.