Practical 2: Matching

Summer 2025
Instructors: Ting Ye & Qingyuan Zhao

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) using data from NHANES I Epidemiologic Follow-up Study.

The dataset nhanesi_class_dataset.csv is posted on the 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 ($\leq$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].

[1]Some NHANES participants were interviewed after 1971 up until 1975. The best way to handle this complication would be to use survival analysis methods with the outcome being time alive since interview and the outcome being censored at the end of follow up. However, since most people were interviewed around 1971 and follow ended for all people in 1992, using years.lived.since.1971.up.to.1992 as the outcome is a good approximation and we will use it for this problem. For discussion of survival analysis methods, see D.R. Cox and D. Oakes, Analysis of Survival Data, Chapman and Hall/CRC, 1984 and P.R. Rosenbaum, Observational Studies, Chapter 2.8.

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. Fit a propensity score model adjusting for the confounders and the two missing indicators in Q1. To find a subset of the units with overlap, follow the procedure of Dehejia and Wahba (1999, Journal of American Statistical Association): exclude from further analysis any treated unit whose propensity score is greater than the maximum propensity score of the control units and exclude any control unit whose propensity score is less than the minimum propensity score of the treated units. How many (if any) units are excluded by this procedure?

3. Form optimal matched pairs using rank based Mahalanobis distance with a propensity score caliper, using the following prognostically important variables in the Mahalanobis distance – smoking status, sex and age at time of interview. Assess the balance on the confounders between the treated and control matched pairs. Compare it with the balance before matching in terms of the *standardized differences* (Lecture 2). Construct a Love plot.

   Remark: The function optmatch_caliper in optmatch.R (posted on the course website) implements the matching. An example of code is as follows:

```
# Specify the model for propensity score (fitted by logistic regression)
ps.formula=physically.inactive~sex+smoking.status+income.poverty.ratio+
age.at.interview+race+education+working.last.three.months+married+
alcohol.consumption+dietary.adequacy+income.poverty.ratio.missingind+
dietary.adequacy.missingind

# Specify all the variables you want to use in the Mahalanobis distance
# on the right handside of ~
mahal.formula=physically.inactive~sex+smoking.status+age.at.interview

# Perform paired matching
# Print standardized difference (both before and after matching) and the Love plot
match_res1<-optmatch_caliper(df,nocontrols.per.match = 1, calipersd=0.5,
ps.formula=ps.formula,mahal.formula=mahal.formula)
match_res1$p
```

4. Using the pair matching from Q3, find a point estimate and 95% confidence interval for the effect of being physically inactive compared to being physically active on years.lived.since.1971.up.to.1992.

   Remark: For example, we can fit a regression model controlling for the matched set indicators, as follows

```
matched.reg.model=lm(years.lived.since.1971.up.to.1992~physically.inactive+matchvec+
sex+smoking.status+income.poverty.ratio+age.at.interview+race+
education+working.last.three.months+married+alcohol.consumption+
dietary.adequacy+income.poverty.ratio.missingind+
dietary.adequacy.missingind,data=df_matched)
```

```
coef(matched.reg.model)[2] # Point estimate of treatment effect
confint(matched.reg.model)[2,] # Confidence interval
```

5. Bonus: Consider matching 2 controls to each treated unit. Is there adequate balance to do so? If there is, consider matching 3 controls to each treated unit and decide if there is adequate balance to do so.

6. Bonus: Which matching that you have considered do you feel is best? Justify your answer.