SPPH 604 001 Exercise: Propensity score matching

26 January, 2024

Contents

Problem 1: [0% grade] **2**

1(a) Importing dataset 2

1(b) Subsetting according to eligibility 2

1(c) Run the design-adjusted logistic regression 3

Problem 2: Propensity score matching by DuGoff et al. (2014) [50% grade] **4**

2(a): 1:1 matching 4

2(b): Interpretation 5

Problem 3: Propensity score matching by Austin et al. (2018) [50% grade] **5**

3(a): 1:4 matching 5

3(b): Interpretation 5

Knit your file **5**

We will use the article by Moon et al. (2021). We will reproduce some results from the article. The authors aggregated 4 NHANES cycles 2005-12 to create their analytic dataset. The full dataset contains 40,790 subjects with the following relevant variables for this exercise:

Survey information

- SEQN: Respondent sequence number
- strata: Masked pseudo strata (strata is nested within PSU)
- psu: Masked pseudo PSU
- survey.weight: Full sample 8 year interview weight divided by 4
- survey.cycle: NHANES cycle

Outcome variable

- cvd: Cardiovascular disease

Exposure

- nocturia: Binary nocturia

Confounders and other variables

- age: Age in years at screening
- gender: Gender
- race: Race/Ethnicity
- smoking: 100+ cigarettes in life
- alcohol: Alcohol consumption (12+ drinks in 1 year)
- sleep: Sleep duration, h
- bmi: Body Mass Index in kg/m^2
- systolic: Systolic blood pressure, mmHg
- diastolic: Diastolic blood pressure, mmHg
- tcholesterol: Total cholesterol, mg/dl
- triglycerides: Triglycerides, mg/dl
- hdl: HDL-cholesterol, mg/dl
- diabetes: Diabetes mellitus
- hypertension: Hypertension

Two important **warnings** before we start:

- In this paper, there is insufficient information to create the analytic dataset. This is mainly because of not sufficiently defining the covariates and not explicitly explaining the inclusion/exclusion criteria.
- The authors did not consider survey features. Since we will utilize survey features in our analysis, our results will likely be different than the results shown by the authors in Table 2.

Problem 1: [0% grade]

1(a) Importing dataset

```
load(file = "Data/propensityscore/Moon2021.RData")
ls()
```

```
## [1] "dat.full"
```

1(b) Subsetting according to eligibility

```
# Age 20+
dat.analytic <- dat.full[complete.cases(dat.full$age),]

# Complete outcome and exposure information
dat.analytic <- dat.analytic[complete.cases(dat.analytic$cvd),]
dat.analytic <- dat.analytic[complete.cases(dat.analytic$nocturia),]

# Keep important variables only
vars <- c(
  # Survey features
  "SEQN", "strata", "psu", "survey.weight",

  # Survey cycle
  "survey.cycle",
```

```
  # Binary exposure
  "nocturia",

  # Outcome
  "cvd",

  # Covariates
  "age", "gender", "race" , "smoking", "alcohol", "sleep", "bmi", "diabetes",
  "hypertension", "tcholesterol", "triglycerides", "hdl", "systolic", "diastolic")

dat.analytic <- dat.analytic[,vars]

# Complete case
dat.analytic <- na.omit(dat.analytic) #  N = 15,404 (numbers do not match with Fig 1)
dim(dat.analytic)
```

```
## [1] 15404    21
```

1(c) Run the design-adjusted logistic regression

Create the first column of Table 2 of the article, i.e., explore the relationship between binary nocturia and CVD among adults aged 20 years and more. Adjust the model for the following covariates: age, gender, race, body mass index, smoking status, alcohol consumption, sleep duration, total cholesterol, triglycerides, HDL-cholesterol, hypertension, diabetes mellitus, and survey cycles.

Note:

- The authors did not utilize the survey features (e.g., strata, psu, survey weights). But you should utilize the survey features to answer this question.
- You must create your design on the full data and then subset the design.
- Report the odds ratio with the 95% CI.

```
# Create an indicator variable in the full data
dat.full$indicator <- 1
dat.full$indicator[dat.full$SEQN %in% dat.analytic$SEQN] <- 0

# Design setup
svy.design0 <- svydesign(strata = ~strata, id = ~psu, weights = ~survey.weight,
                         data = dat.full, nest = TRUE)

# Subset the design
svy.design <- subset(svy.design0, indicator == 0)

# Design-adjusted logistic
fit.logit <- svyglm(I(cvd == "Yes") ~ nocturia + age + gender + race + bmi +
                      smoking + alcohol + sleep + tcholesterol + triglycerides +
                      hdl + hypertension + diabetes + survey.cycle,
                    family = binomial, design = svy.design)
publish(fit.logit)
```

```
##       Variable              Units OddsRatio         CI.95     p-value
##       nocturia                 <2       Ref
##                                2+      1.44   [1.21;1.71]   0.0001496
##            age            [20,40)       Ref
##                           [40,60)      4.21   [3.05;5.82]     < 1e-04
##                           [60,80)     11.46  [7.89;16.64]     < 1e-04
##                          [80,Inf)     25.28 [17.51;36.50]     < 1e-04
##         gender               Male       Ref
##                            Female      0.68   [0.58;0.79]     < 1e-04
##           race          Hispanics       Ref
##                Non-Hispanic White      1.32   [1.10;1.57]   0.0036168
##                Non-Hispanic Black      1.15   [0.92;1.44]   0.2362499
##                       Other races      1.55   [1.05;2.30]   0.0319116
##            bmi                         1.02   [1.01;1.03]   0.0003273
##        smoking                 No       Ref
##                                Yes      1.74   [1.46;2.07]     < 1e-04
##        alcohol                 No       Ref
##                                Yes      0.92   [0.59;1.45]   0.7273627
##          sleep                          0.96   [0.90;1.01]   0.1146287
##   tcholesterol                          0.99   [0.99;0.99]     < 1e-04
##  triglycerides                          1.00   [1.00;1.00]   0.4801803
##            hdl                          0.99   [0.98;1.00]   0.0416900
##   hypertension                 No       Ref
##                                Yes      2.73   [2.27;3.29]     < 1e-04
##       diabetes                 No       Ref
##                                Yes      1.83   [1.51;2.22]     < 1e-04
##   survey.cycle            2005-06       Ref
##                           2007-08      0.84   [0.65;1.07]   0.1644272
##                           2009-10      0.91   [0.73;1.12]   0.3793696
##                           2011-11      0.82   [0.68;0.99]   0.0398975
```

Problem 2: Propensity score matching by DuGoff et al. (2014) [50% grade]

2(a): 1:1 matching

Create the second column of Table 2 (exploring the relationship between binary nocturia and CVD; the same exposure and outcome used in Problem 1) using the propensity score **1:1 matching** analysis as per DuGoff et al. (2014) recommendations.

You should consider all four steps in the propensity score (PS) analysis:

- Step 1: Fit the PS model by considering survey features as covariates. Other covariates for the PS model are the covariates used in 1(c).
- Step 2: Match an exposed subject (nocturia $\geq$ 2 times) with a control subject (nocturia <2 times) without replacement within the caliper of 0.2 times the standard deviation of the logit of PS. Set your seed to 123.
- Step 3: Balance checking using SMD. Consider SMD <0.1 as a good covariate balancing.
- Step 4: Fit the outcome model on the matched data. If needed, adjust for imbalanced covariates in the outcome model. Report the odds ratio with the 95% CI. You should utilize the survey feature as the design (NOT covariates).

```
# your codes here
```

2(b): Interpretation

Compare your results with the results reported by the authors. [Expected answer: 1-2 sentences]

Problem 3: Propensity score matching by Austin et al. (2018) [50% grade]

3(a): 1:4 matching

Repeat Problem 2(a), i.e., create the second column of Table 2 (exploring the relationship between binary nocturia and CVD), but using the propensity score **1:4 matching** analysis as per Austin et al. (2018) recommendations.

You should consider all four steps in the propensity score (PS) analysis:

- Step 1: Fit the PS model by considering survey features as design, i.e., fit the design-adjusted PS model. Other covariates for the PS model are the covariates used in 1(c).
- Step 2: Match an exposed subject (nocturia $\geq$ 2 times) with 4 control subjects (nocturia <2 times) with replacement within the caliper of 0.2 times the standard deviation of the logit of PS. Set your seed to 123.
- Step 3: Balance checking using SMD. Consider SMD <0.1 as a good covariate balancing. Remember, you need to consider matching weights in checking the covariate balance.
- Step 4: Fit the outcome model on the matched data. If needed, adjust for imbalanced covariates in the outcome model. Report the odds ratio with the 95% CI. You should utilize the survey feature as the design (NOT covariates).

Note:

- For step 4, you need to multiply matching weights and survey weights when creating your design. After creating the design with the new weight, subset the design for the matched sample. This step is required to get survey-based estimates.

```
# your codes here
```

3(b): Interpretation

Compare the results with Problem 2. What's the overall conclusion? [Expected answer: 2-3 sentences]

Knit your file

Please knit your file once you finished and submit the knitted PDF file **ONLY**. Please also fill-up the following table:

Group name: ** xyz **

Student initial	% contribution
Student 1 initial	x%
Student 2 initial	x%
Student 3 initial	x%