SPPH 604 001 Lab Exercise: Survey data analysis

25 January, 2024

Contents

Problem Statement

We will use the article. We will use the following article by Flegal et al. (2016).

We will reproduce some results from the article. The authors used NHANES 2013-14 dataset to create their main analytic dataset. The dataset contains 10,175 subjects with 12 relevant variables:

- SEQN: Respondent sequence number
- RIDAGEYR: Age in years at screening
- RIAGENDR: Gender
- DMDEDUC2: Education level
- RIDRETH3: Race/ethnicity
- RIDEXPRG: Pregnancy status at exam
- WTINT2YR: Full sample 2 year weights
- SDMVPSU: Masked variance pseudo-PSU
- SDMVSTRA: Masked variance pseudo-stratum
- BMXBMI: Body mass index in kg/m**2
- SMQ020: Whether smoked at least 100 cigarettes in life
- SMQ040: Current status of smoking (Do you now smoke cigarettes?)

Question 1: Creating data and table

1(a) Importing dataset

```
# you have to download the data in the same folder
load("Data/surveydata/Flegal2016.RData")
ls()
```

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

```
names(dat.full)
```

```
##  [1] "SEQN"     "RIDAGEYR" "RIAGENDR" "DMDEDUC2" "RIDRETH3" "RIDEXPRG"
##  [7] "WTINT2YR" "SDMVPSU"  "SDMVSTRA" "BMXBMI"   "SMQ020"   "SMQ040"
```

1(b) Subsetting according to eligibility

Subset the dataset according to the eligibility criteria described in the second paragraph of the `Methods` section.

- Hint: The authors restricted their study to
  - adults aged 20 years and more,
  - non-missing body mass index, and
  - non-pregnant.

Your analytic sample size should be 5,455, as described in the first sentence in the `Results` section.

```
# 20+
dat.analytic <- subset(dat.full, RIDAGEYR>=20) # N = 5,769

# Non-missing outcome
dat.analytic <- subset(dat.analytic, !is.na(BMXBMI)) # N = 5,520

# Non-pregnant
dat.analytic <- subset(dat.analytic, is.na(RIDEXPRG) | RIDEXPRG !=
                         "Yes, positive lab pregnancy test") # N = 5,455

dim(dat.analytic)
```

```
## [1] 5455   12
```

1(c) Reproduce Table 1

Reproduce Table 1 of the article.

- Hint 1: The authors reported unweighted frequencies, and thus, survey features should not be utilized to answer this question. Please be advised to order the categories as shown in the table. `tableone` package could be helpful.

- Hint 2: the authors did not show the results for the `Other` race category. But in your table, you could include all race categories.

```
library(tableone)

dat <- dat.analytic

# Age
dat$age <- cut(dat$RIDAGEYR, c(20, 40, 60, Inf), right = FALSE)

# Gender
dat$gender <- dat$RIAGENDR

# Race/Hispanic origin group
dat$race <- dat$RIDRETH3
dat$race <- car::recode(dat$race, " 'Non-Hispanic White'='White'; 'Non-Hispanic Black'=
                        'Black'; 'Non-Hispanic Asian'='Asian'; c('Mexican American',
                        'Other Hispanic')='Hispanic'; 'Other Race - Including Multi-Rac'=
                        'Other'; else=NA", levels = c("White", "Black", "Asian",
                                                      "Hispanic", "Other"))

# Table 1: Overall
tab11 <- CreateTableOne(vars = "age", strata = "race", data = dat, test = F,
                        addOverall = T)

# Table 1: Male
tab12 <- CreateTableOne(vars = "age", strata = "race", test = F, addOverall = T,
                        data = subset(dat, gender == "Male"))

# Table 1: Female
tab13 <- CreateTableOne(vars = "age", strata = "race", test = F, addOverall = T,
                        data = subset(dat, gender == "Female"))

# Reproducing Table 1
tab1a <- list(Overall = tab11, Male = tab12, Female = tab13)
print(tab1a, format = "f") # Showing only frequencies
```

```
## $Overall
##              Stratified by race
##               Overall White Black Asian Hispanic Other
##   n           5455    2343  1115  623   1214     160
##   age
##      [20,40)  1810     734   362  216    412      86
##      [40,60)  1896     759   383  251    449      54
##      [60,Inf) 1749     850   370  156    353      20
##
## $Male
##              Stratified by race
##               Overall White Black Asian Hispanic Other
##   n           2638    1130  556   300   573      79
##   age
##      [20,40)   909     386  182   106   189      46
##      [40,60)   897     360  179   120   215      23
##      [60,Inf)  832     384  195    74   169      10
```

```
##
## $Female
##              Stratified by race
##               Overall White Black Asian Hispanic Other
##   n           2817    1213  559   323   641      81
##   age
##      [20,40)   901     348  180   110   223      40
##      [40,60)   999     399  204   131   234      31
##      [60,Inf)  917     466  175    82   184      10
```

Question 2

2(a) Reproduce Table 1 with survey features [15% grade]

Not in this article but in many other articles, you would see `n` comes from the analytic sample and `%` comes from the survey design that accounts for survey features such as strata, clusters and survey weights. In Question 1, you see how n comes from the analytic sample. Your task for Question 2(a) is to create % part of the Table 1 with survey features, i.e., % should come from the survey design that accounts for strata, clusters and survey weights. You do not need to show the frequiencis but show only the percentages (for categorical variables).

Hints:

- Subset the design, not the sample. For this step, you need to work with your full data. If you have generated a variable in your analytic dataset, that variable should also be present in the full dataset.
- Generate age, gender, and race variable in your full data. Codes shown in Question 1 could be helpful.
- Make the design on the full data and then subset the design.
- Reproduce Table 1 with the design from the previous step. The `svyCreateTableOne` function could be a helpful function.

```
## Create all variables in the full data
# Age
dat.full$age <- cut(dat.full$RIDAGEYR, c(20, 40, 60, Inf), right = FALSE)

# Gender
dat.full$gender <- dat.full$RIAGENDR

# Race/Hispanic origin group
dat.full$race <- dat.full$RIDRETH3
dat.full$race <- car::recode(dat.full$race, " 'Non-Hispanic White'='White';
                             'Non-Hispanic Black'='Black'; 'Non-Hispanic Asian'='Asian';
                             c('Mexican American','Other Hispanic')='Hispanic';
                             'Other Race - Including Multi-Rac'='Other';
                             else=NA", levels = c("White", "Black", "Asian",
                                                  "Hispanic", "Other"))

## Subset the design
# your codes here
```

```
## Table 1
# your codes here

#print(tab1b, format = "p") # Showing only percentages
```

2(b) Reproduce Table 3 [50% grade]

Reproduce the first column of Table 3 of the article (i.e., among men, explore the relationship between obesity and four predictors shown in the table).

- If necessary, re-level or re-order the levels.
- You need to generate obesity as BMI ≥ 30 kg/m^2
- You need to generate `smoking status` and `education`. The unweighted frequencies should be matched with the frequencies in eTable 1 and eTable 2. Make sure these variables are in your full dataset as well.
- Subset the design, not the sample.
- Fit the model. Do not need to report the model summary.
- The authors used SAS to produce the results vs. We are using R. The estimates could be slightly different (in second decimal point) from the estimates presented in Table 3, but they should be approximately similar.
- You can use `Publish` or `jtools` package to report the odds ratios. Your odds ratios could be look like as follows:

Variable	Units	OddsRatio
age	[20,40)	Ref
	[40,60)	1.28
	[60,Inf)	1.20
race	White	Ref
	Black	1.24
	Asian	0.27
	Hispanic	1.22
	Other	1.23
smoking	Never smoker	Ref
	Former smoker	1.25
	Current smoker	0.71
education	High school	Ref
	<High school	0.93
	>High school	0.97

```
# your codes here
```

2(c) Model selection [25% grade]

From the literature, you know that `age` and `race` needs to be adjusted in the model, but you are not sure about `smoking` and `education`. Run an AIC based backward selection process to figure out whether you want to add `smoking` or `education`, or both in the final model in 2(b). What is your conclusion, i.e., which variables are selected/dropped [Expected answer: one short sentence]?

Hints:

- Your design must be free from missing values. Even after applying eligibility criteria, you may have some missing values on multiple variables (see eTable 1 and eTable 2). This is especially important for model selection process.
- You can create a complete case analytic dataset (i.e., dataset without missing values in obesity, four predictors, and survey features). Then create the design on the full data and subset the design for the complete case samples.
- `step` function could be helpful.

```
# your codes here
```

2(d) Testing for interactions [10% grade]

Check whether the interaction between `age` and `smoking` should be added in the 2(b) model (yes or no answer required, along with the code and p-value):

```
# your codes here
```

Knit your file

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

Group name: **Put the group name here**

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