Exercises

Thomas Lumley

24/07/2019

Session 1

(A) The PISA educational survey is stratified by country and then by various factors within country. A sample of schools is taken in each stratum, and students are sampled from each school.

`nzmaths.csv` is the New Zealand subset of some variables related to mathematics performance.
`nzmaths.pdf` is documentation.

Declare a survey design object for these data. Use `svytotal` to estimate the total number of male and female school students at the survey age in New Zealand.

(B) The data sets `esophlong.csv` is the famous oesophageal cancer case-control study of Tuyns and co-workers, used by Breslow and Day. A case-control study is a stratified sample, stratified on case status. Based on the population size for the region being studied, approximately 1/440 controls were sampled. Declare a survey design for these data

(C) The file `pfoa-nhanes.rda` contains a set of data frames (also in the files `pfoa00.dta` to `pfoa10.dta`). Read in `pfoa00.dta`.

Declare a survey design using the strata (`SDMVSTRA`), the sampling units (`SDMVPSU`) and the weights for the medical and clinical examination (`WTMEC2YR`). Estimate the population total number of men and women (`RIAGENDR`) and each race/ethnicity group (`RIDRETH1`)

Now repeat this using the replicate weights `WTMREP01` to `WTMREP52` instead of strata and cluster information.

Session 2

With the `pfoa00` data set, estimate the mean age, and the mean and quartiles of `SPFOA` (perfluorooctanoic acid in blood, ng/ml)

With the PISA maths data, draw a histogram of `PCGIRLS` , the proportion of girls at the school. Draw a scatterplot against `PV1MATH` (maths score) and against `MATHEFF` (maths self-efficacy score)

Fit an unweighted logistic regression model to the `esophlong` data set with linear effects of alcohol and tobacco and age. Compare to a model treating it as a survey sample, fitted using `svyglm` or `svy: logistic`

Session 3

We are interested in whether exposure to perfluorooctanoic acid (PFOA) is associated with cardiovascular events (CVD) or with peripheral vascular disease (PAD)

In R, the lines of code

```
nhanes00<-svydesign(id=~SDMVPSU,strata=~SDMVSTRA,weights=~WTMEC2YR,data=alldata00,nes
t=TRUE)

nhanes00<-update(nhanes00, hadcvd=(MCQ160C==1) | (MCQ160E==1) | (MCQ160F==1), abi=(LE
XRABPI+LEXLABPI)/2)
nhanes00<-update(nhanes00, haspad=ifelse(abi<0.9,1,ifelse(abi>1.5,NA,0)))
svyquantile(~SPFOA,nhanes00,quantiles=c(0.25,0.5,0.75))
nhanes00<-update(nhanes00, pfoa4=cut(SPFOA,c(0,3.7,5,6.8,Inf)))
nhanes00<-update(nhanes00, smoking=ifelse(SMQ020==2,0,ifelse(SMQ040 %in% c(1,2),1,
2)))
```

set up the survey design. In Stata, the variable declarations can just be done as usual with `gen` or `replace` after `svyset`.

Try logistic models for `hadcvd` or `haspad`, with `pfoa4` as a predictor. In addition to gender and race/ethnicity (as in session 1) consider adjustment variables:

name	**variable**
BPXSAR	systolic blood pressure
BPXDAR	diastolic blood pressure
BMXBMI	BMI
LBXTC	total cholesterol
LBXGH	% glycosylated hemoglobin
smoking	smoking
DMDEDUC	education
RIDAGEYR	age

If you have time, try `pfoa4`

```
nhanes04<-svydesign(id=~SDMVPSU,strata=~SDMVSTRA,weights=~WTSA2YR,data=alldata04,nest
=TRUE)


nhanes04<-update(nhanes04, hadcvd=(MCQ160C==1) | (MCQ160E==1) | (MCQ160F==1), abi=(LE
XRABPI+LEXLABPI)/2)
nhanes04<-update(nhanes04, haspad=ifelse(abi<0.85,1,ifelse(abi>1.5,NA,0)))
svyquantile(~LBXPFOA,nhanes04,quantiles=c(0.25,0.5,0.75),na.rm=TRUE)
nhanes04<-update(nhanes04, pfoa4=cut(LBXPFOA,c(0,2.8,4.2,6,Inf)))
nhanes04<-update(nhanes04, smoking=ifelse(SMQ020==2,0,ifelse(SMQ040 %in% c(1,2),1,
2)))
```

or `pfoa10`

```
nhanes10<-svydesign(id=~SDMVPSU,strata=~SDMVSTRA,weights=~WTSC2YR,data=alldata10,nest
=TRUE)


nhanes10<-update(nhanes10, hadcvd=(MCQ160C==1) | (MCQ160E==1) | (MCQ160F==1))
svyquantile(~LBXPFOA,nhanes10,quantiles=c(0.25,0.5,0.75),na.rm=TRUE)
nhanes10<-update(nhanes10, pfoa4=cut(LBXPFOA,c(0,2.6,4.2,6.2,Inf)))
nhanes10<-update(nhanes10, smoking=ifelse(SMQ020==2,0,ifelse(SMQ040 %in% c(1,2),1,
2)))
```

with adjustment variables

name	variable
BPXSY1	systolic blood pressure
BPXDI1	diastolic blood pressure
BMXBMI	BMI
LBXTC	total cholesterol
LBXGH	% glycosylated hemoglobin
smoking	smoking
DMDEDUC2	education
RIDAGEEX	age