week 8 - Class/Take Home exercise nonresponse weighting

Peter Lugtig

16 oktober, 2023

In this exercise, you will use a condensed version of the European Social Survey (ESS) Round 5 conducted around 2010. The dataset is condensed in two ways: 1. It only includes Hungary and Slovakia out of a total of 32 countries 2. It contains a small subset of the variables asked in the survey

The ESS is a cross-national survey with data collection being conducted face-to-face. The sampling design varies slightly per country but in every country PSUs are drawn (clusters), and in every PSU a subset of individuals is selected. Some countries use stratification, either to select the PSUs (e.g. larger PSUs with larger probabilities), or select cases within PSUs. Lets load the data:

```
data <- readRDS("ESSR5_HUSK.RDS")
```

The dataset contains several variables that describe the sampling design. The way that ESS round 5 has done this is somewhat different than other rounds. The following variables are available:

- PSU: number indicating the PSU
- CNTRY: the country (Hungary or Slovakia).
- PROB: The inclusion probability of individuals within every PSU.
- STRATIFY: the stratifcation variable. Note that this is a regional indicator in the case of ESS. We will for now just use the PSUs

For the first questions, we will just use Slovakia, so in a first step, we will select just those cases:

```
SK <- subset(data,CNTRY=="SK")
```

Question 1: The survey design object

Although the ESS contains information on the cluster, it does not contain information on the size of the clusters, so we cannot specify the survey design like we did in the first weeks of this course. We have to rely on a variable included in the ESS "dweight" that is design weight: a (normalized) inverse probability of every individual to be selected in the sample. This implies we will use the Hurvitz-Thompson estimator in this whole exercise (see week 6).

Specify the survey design object for Slovakia, using both the "PSU" variable to indicate clustering, and the "dweight" variable as the weight variable. Ignore nonresponse for now. Please note that the "dweight" variable sums up to the sample size. In earlier practicals, we always included an fpc-correction. If you want to do this correctly, multiply the dweight by 4261000/samplesize(SK) = {4261000/nrow(SK)}

(adult population size Slovakia = 4.3 million).

(short intermezzo - shouldn't we specify the fpc at the cluster level as in {id=~PSU+id}? The practical answer is that we have a very large population size, and so fpc won't matter much. But the bigger problem is that we want to have the population size at the level of the psu, which we don't have. This is a common problem; if we sample regions, we often don't know the exact population size. Using sample weights that add up to the population size, under HT-estimators at least correct for the overall population size.)

Question 2: Daily Internet use (y)

For the exercise of today, we will try to explain variation in Internet Use (in 2010) in Slovakia. The variable Internet Use (NETUSE) is an ordinal variable, but we will use the variable DAILYINT, which is a binary variable indicating whether people use the Internet daily (1) or not (0). What is the proportion of daily Internet Users in Slovakia, ignoring nonresponse, and what is the confidence interval around this estimate? (svyciprop is a handy function here).

Question 3 (optional): Understanding values of design weights

What does the Design weight actually mean? The ESS contains three indicators that together tell us how the survey was designed. Use the variables "PROB", "dweight" and "hhmmbr" (the number of household members in a house including children) to explore how the design weight relates to the cluster (PSU) and number of household members sampled.

Question 4: Predictors of nonresponse

The ESS also includes Nonresponse weights, that correct for nonresponse using poststratification to the three-way crosstable of gender, age and educational level, as well as the sampling design. Can you plot the relation between "pspwght", "gndr", "agea" and "eduyrs"? Which variable(s) are a strong predictor of nonresponse in the ESS-Slovakia?

Question 5: Correcting for nonresponse

Estimate the proportion of daily internet users again, but now correct both for sampling design and nonresponse differences by using the variable "pspwght". Is your conclusion different than in question 2?

Question 6: Why a difference?

If everything went correctly, you just found that the proportion of Internet Users in Slovakia changes after you corrected for nonresponse. Can you explain why this is the case when you look at the relation between "DAILYINT" and the variables that are used in creating the nonresponse weights?

Question 7: What missing data mechanism?

Based on your findings in question 6, what would you say is your conclusion about the nonresponse mechanism in the ESS-Slovakia? Is there:
- Missing Completely at Random?
- Missing at Random?
- Missing Not at Random?

Question 8: Why the change in s.e. after nonresponse correction?

Next week, we will talk in quite a bit of detail about how to design weights. Perhaps you noticed in question 6 that the standard error of our estimate becomes quite a bit larger when we weight for sampling and nonresponse. The great thing about weighting is that it can be used to correct for bias. This sometimes comes at the price of an increased variance. Can you investigate what the difference is between the highest and lowest pspwgt in the dataset?

Question 9: What happens when we fit a regression model?

Some people (e.g. economists) argue that survey weighting is perhaps useful when you are estimating a point estimate like the proportion of daily Internet Users, but not so much when you are actually interested in a regression model.

Can you try to explain variation in Daily Internet Use using predictors in the ESS ? It doesn't really matter what predictors you use (just take a few demographic variables, like gender, age, income, level of education).

After selecting predictors from the dataset, run the model 3 times:
1. Run the model first using only covariates, and no sampling or nonresponse weights 2. a second time using design weights (use the right svydesign object and the svyglm function), and then 3. a third model correcting both for sampling design differences and nonresponse.

Do the regression coefficients change? (note that the variable netuse also measures Internet use, so do not use this is a predictor)

Question 10 (optional):

In the week before the break, we talked about regression modeling as a way to do inference. Let's here focus again on estimating the number of daily Internet users in Slovakia. Can we also use a regression-estimation model here? Yes, we can! What we need is:
1. A *good* regression model for explaining our dependent variable.
2. Means (or proportions) for our covariates to plugin to our regression model. Rather than using the raw means, we can here use the weighted means now for our covariate (or use statistics from external sources - more on this next week)

The code below shows how regression modeling would handle this issue. Run the code, and inspect the output. Do you find differences when you compare your results to question 5 where we also corrected for nonresponse?

```
# step 1, a good regression model with weighted covariates.
# we will just use a simple here model from question 6 here.
# ideally, you want this model to be better, but just to illustrate the approach
# we use a small set. link = "logit" indicates we want a logistic model.
# you can ignore the warning here.

wresults <-svyglm(DAILYINT ~ agea+gndr+eduyrs,design=SKdesignNR,
                  family=binomial(link="logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
#step 2: add the new predicted values:
# add predicted values for "dailyint" to a new dataset
# the Type="response"  indicates we want to predict on the scale of y (binary)
# We thus get the "predicted probabilities" of using Internet daily, also known
# as response propensities
# note: first create a new dataset
newdata <- subset(SK, select=c(agea,eduyrs,gndr,DAILYINT))
# we have missing data however in our predictors!
# we will cover how to deal with these in two weeks, when we discuss imputation.
# For now, I just select the complete cases, which means that I delete about 40 rows.
newdata <- newdata[complete.cases(newdata),] # selecting only complete rows

newdata$DAILYINTwgtpred <- predict(wresults,newdata=newdata,type="response") #weighted
mean(newdata$DAILYINTwgtpred,na.rm=T) # .33 check
```

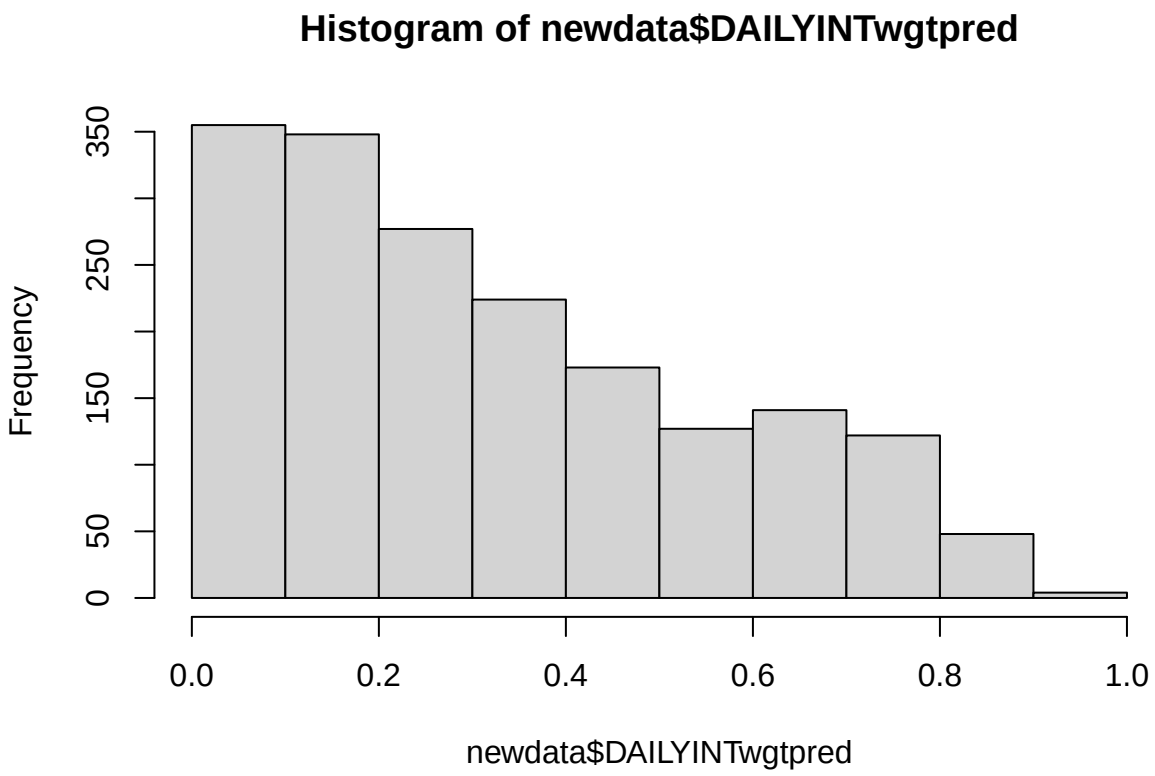
```
## [1] 0.3251904
```

```
hist(newdata$DAILYINTwgtpred,na.rm=T) # these are response propensities
```

```
## Warning in plot.window(xlim, ylim, "", ...): "na.rm" is not a graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "na.rm"
```

```
## is not a graphical parameter
## Warning in axis(1, ...): "na.rm" is not a graphical parameter
## Warning in axis(2, ...): "na.rm" is not a graphical parameter
```



So far we have only used data from Slovakia. Sometimes we want to combine data from multiple populations. For example, imagine we want to use data from the European Social Survey to estimate daily internet use in the European Union. In our case, we want to estimate statistics for Hungary and Slovakia combined. We now face the problem, that sample sizes across the countries are similar, but the population sizes between Slovakia and Hungary are very different.

The variable "pweight" is a population size equivalence weight: it weights data from the different subpopulations when you combine them in one analysis.

(optional) Question 11: population size equivalence weights

Using the variable "pweight", can you estimate the population size in Hungary, knowing that Slovakia's population is about 4.2 million?

Question 12: Creating a new weight to do analysis including multiple countries

Before we go and estimate daily Internet use in both countries, we have to create a new weight in our original dataset. We have to multiply the "pspwght" with the population equivalence weight "pwght" to get a new analysis weight "anweight". Create this new analysis weight to the original dataset including both countries

Question 13:

Please note that the pweight is not centered at 1, and so we again do not take into account the fpc-correction. However, as the population is so large, this is not a big problem, and we will ignore fpc. Estimate the proportion of daily internet users in both countries. Use the svyby() command to do this (and look up how this works).

Question 14 (optional): why the difference?

Did you encounter a difference between the proportion of Daily Internet users in Slovakia between question 13 and your earlier question? You should find a small difference, which is strange, because you are asking the same thing as before, except you have now weighted all cases in Slovakia with the same number. If you have time left, investigate why this may be (hint: missing data)

– End of document –