**Statistical Methods in Epidemiology**

**Lab 4-5**

**Analysis of surveys with complex sampling design**

We will use data from the EMENO study, a cross-sectional health examination survey conducted in Greece in 2014-2016. The EMENO utilized a multistage stratified random sample design to select a sample of 6,006 adults (≥18 years) living in Greece based on the 2011 census. At the first stage of sampling, all prefectures in the country were stratified by geographical region and by the degree of urbanization into 33 strata, and two prefectures per stratum were randomly selected. Sampling fractions were allowed to differ by stratum over-representing under-populated strata and under-representing over-populated strata. In the second stage of sampling, a sample of area segments (sampling points) was randomly selected (without replacement). In the third stage of sampling, a systematic sample of eligible households was selected from each segment with sampling intervals of two to four depending on population density. Finally, one adult individual was randomly selected from each sampled household (the one who had the most recent birthday). Thus, the sampling weights were the inverse of sampling fractions (or inclusion probability).

1. **Construction of the weights**
2. Let’s have a look at the dataset “analysis.dta”, where the design variables plus gender and age are included.

desc

1. The variable base\_weight corresponds to the sampling weight. Look at their distribution. What do you observe? How can we interpret the sampling weights?
2. When we analyze survey data it is necessary to specify the design variables (PSUs, strata, fractions) to have representative data. In Stata, these variables are specified once using the svyset command and then they are used for all subsequent survey analysis svy: commands.

 The syntax of the svyset command is as follows:

svyset PSU [pweight=…], strata()

In our example, the primary sampling units (PSU’s) are the sampling points (variable blockid), and the probability weights (pweight) are (for now) the sampling weights (base\_weight). Thus, we type:

svyset psu [pweight=base\_weight], strata(strata)

When there is only one PSU within a stratum, there is insufficient information to compute an estimate of that stratum's variance. Therefore, it is impossible to compute the variance of an estimated parameter when the data are from a stratified clustered design. To overcome this issue, we can add in the svyset command the option “singleunit(centered)” which specifies that strata with singleton PSUs will be centered at the grand mean instead of the stratum mean.

svyset psu [pweight=base\_weight], strata(strata) singleunit(centered)

 Once the data have been svyset we can use any of the svy commands without having to specify the design variables of the survey each time. Before every command, we add the prefix “svy:”. The majority of the commonly used commands (logit, regress, tab, etc.), are compatible with svy. E.g., if we want to estimate the age group distribution of the population we type:

svy: tab age\_gr

We can add the options “per” and “obs” to show the percentages (instead of proportions) and the number of observations respectively.

svy: tab age\_gr, per obs

How can we estimate the area distribution of the sample and the respective one of population? Why do we observe such large differences?

1. The table below shows the age and gender distribution as provided by ELSTAT.

Run the same analysis in the dataset. Are the results close?

|  |  |  |
| --- | --- | --- |
|  | Πανελλαδικά |  |
|  | Πληθυσμός % |  |
| Ηλικία  | Άνδρας | Γυναίκα |
| 18-29  | 9.0 | 8.5 |
| 30-39  | 9.2 | 9.0 |
| 40-49  | 8.8 | 9.0 |
| 50-59  | 7.6 | 8.0 |
| 60-69  | 6.1 | 6.6 |
| 70-79  | 5.8 | 6.2 |
| 80+  | 2.7 | 3.9 |
| Σύνολο  | 48.6 | 51.4 |

1. Το adjust for those discrepancies, we need the age and gender distribution by geographical region from the 2011 census (file census). We need to construct the so-called post-stratification factor to obtain the final weights. Let’s first have a look at the file “census”.

|  |  |  |  |
| --- | --- | --- | --- |
| gender | age\_gr | area | freq |
| 0 | 0 | 1 | 296461 |
| 0 | 1 | 1 | 317576 |
| 0 | 2 | 1 | 282668 |
| 0 | 3 | 1 | 235520 |
| 0 | 4 | 1 | 182732 |

Frequencies, by area, age\_gr and gender are provided. We need to calculate factors that will be multiplied with the initial sampling weights to provide the above frequencies. Note that, from now on, all the observations with missing age or gender will be excluded.

1. We now have constructed the final weights to provide estimates representative of the population. Run svy:tab age\_gr gender, per obs to see the age and gender distribution after having adjusted the weights to match the distribution of the population. The results should be close to those of the table in question 4 (the age and gender distribution as provided by the ELSTAT).
2. **Main analysis**
3. Let’s proceed with the main analysis. We want to estimate the prevalence of several CVD risk factors in Greece (diabetes, hypertension, hypercholesterolemia) and assess the associated factors. We merge our file with the “data\_exams.dta” file.

 merge 1:1 emenoid using "G:\works\SME\data\_exams"

 Result # of obs.

 -----------------------------------------

 not matched 1,695

 from master 1,695 (\_merge==1)

 from using 0 (\_merge==2)

 matched 4,298 (\_merge==3)

We notice that 1695 individuals do not exist in the exams dataset. This is actually a missing data problem, arising very often in such studies because not all of the participants consent to physical examination or to provide a blood sample. Intuitively, we understand that if people that provide a blood sample are different in some characteristics from those who do not, our estimates will be biased. For example, if we want to estimate hypercholesterolemia’s prevalence, and younger individuals tend not to participate in the examination phase of the survey, then hypercholesterolemia’s prevalence will be overestimated. How can we investigate if there are differences between these two subsamples?

1. Look at the merged dataset and give some descriptives to see how svy commands work. The following commands can be used:

svy:tab gender, per obs

svy:mean age

estat sd

1. Estimate the prevalence of the variables of interest (diabetes, hypertension, hypercholesterolemia).

svy:tab diab\_prev, per obs ci

svy: tab hypertension\_pr, per obs ci

svy:tab hyperchol200, per obs ci

The prevalence is usually accompanied by the respective 95% CI.

1. Is hypertension associated with BMI? Check for potential confounders.
2. Estimate the prevalence of those having ldl>130 mg/dL (variable hyperldl130). Does the prevalence of high ldl differ between males and females? Run a weighted logistic model to assess the factors associated with high ldl (try age, gender and bmi).
3. Interpret the results of the previous model. Try the following commands to plot the predicted probability of high ldl by gender over age:

margins gender, at(age=(18(10)100)) vsquish

marginsplot

1. Now, let’s investigate if the relationship between age and the logit of the probability of high ldl is linear, or we need to include some non linear terms in the model. How can we investigate what is the appropriate functional form of age? After addressing that question, run the final model.
2. Continuous by continuous interactions: Let’s investigate the factors associated with hypercholesterolemia (chol>200), but now we will include bmi as continuous rather than categorical, and its interaction with age. Other significant factors are gender, and walking>30 minutes a day. Run a model including all the above predictors and try to interpret the results.