Exercise 4.1: Sampling and analysing a two-stage study

Data from a two-stage study may include all study subjects (i.e. the first stage) in the data set, but the covariate that is only available for the second stage subjects has a missing value for the others. Thus, the proportion of second-stage individuals in the various strata can be readily found from the counts of the total and the complete records in the strata. The inverse of this "sampling fraction" provides the weight for the weighted logistic regression. While these operations can be easily done with standard commands in statistical software, user-written packages are also available: for example, the **meanscor** command in Stata and the **survey** package in R.

Two-stage data may also be available in two datasets, one with the first stage data (i.e. all the study subjects with their first stage variables) and another separate dataset for the second stage subjects. If there is a common identifier, the datasets can be merged to create the same structure just described. If there is no common identifier, the first stage dataset can be aggregated to get the counts in all the strata, merged (by stratum) to the second stage data. The weights are simply the ratios of the first and second stage sample sizes in each of the strata.

1. In this exercise, you will sample a two-stage study from a cross-sectional sample of individuals from the Framingham cohort study, whose data are in the dataset **Framingham_HDL.dta** provided. This data consists of 3,026 cohort members who had HDL and LDL measurements taken at their third visit. The outcome of interest is any CHD, with 777 cases and 2249 non-cases. Our objective is to estimate the effect of HDL and LDL on any CHD, adjusted for sex and hypertension. The variables are:

randid (individual identification number) **anychd** (case indicator: 0=no CHD during follow-up, 1=CHD during follow-up) **sex** (1=male, 2=female) **cursmoke** (0=not current smoker, 1=current smoker) **prevhyp** (0=no prevalent hypertension, 1=prevalent hypertension) **ldlc** (low-density lipoprotein cholesterol, mg/dL) **hdlc** (high-density lipoprotein cholesterol, mg/dL) **ldlc_cat** (categorical LDL: desirable = ≤100mg/dL, moderate = 101-159mg/dL, high = 160-189mg/dL, very high = ≥190mg/dL) **hdlc_cat** (categorical HDL: low = ≤40mg/dL, borderline = 41-59mg/dL, desirable = ≥60mg/dL) **hdlc_cat2** (**hdlc_cat**, releveled so 'desirable' is the reference)

i. Run a logistic regression on the full cross-sectional sample with **anychd** as outcome and covariates **ldlc_cat, hdlc_cat2, sex**, and **prevhyp**, to get adjusted ORs for HDL and LDL.

ii. Now we will mimic a two-stage study where HDL and LDL were considered as 'expensive' covariates and were only measured for a second-stage subsample and were missing for the other individuals. To mimic a situation where the second-stage sample consists of all 777 cases and an equal number of controls, select/identify a subsample of 777 controls (using a random number generator) and save a dataset containing all the cases and these controls (*Note. You should set a "seed" so that you can reproduce the exact same data if needed*).

iii. For the dataset selected in (ii), create a weight variable that is 1 for all cases and 2249/777 for each of the controls. Verify that running a weighted logistic regression (with robust SE) of *just these (second-stage) data* gives similar estimates to the full cohort results in (i).

iv. Return to the full data set, and this time when you select the 777 controls, do not delete the others from the data, but set their HDL and LDL variables to missing, and add a second-stage indicator variable to the data which is 0 for these missing observations and 1 for the cases and 777 controls. Repeat the logistic regression analysis this data set using the **meanscor** command in Stata or the **survey** package in R and compare your results to what you obtained with weighted regression in (iii).

2. This exercise uses some of the data collected in a cross-sectional survey of schoolchildren in Stockholm schools to determine the risk of H. Pylori infection in children from infected family members. For all study children, information was gathered on household socioeconomic status (SES), and for immigrant children, whether the family's country of origin had a high or low H. Pylori prevalence. The information on H. Pylori infection in family members was only collected for a sub-sample of children. The family members of all infected children were tested, but in an effort to target the control families considered most informative, the intensity of sampling was different in the strata defined by SES (high vs. Low) and the H. Pylori prevalence (high vs. Low) in the child's country of origin. Thus the second-stage sampling is stratified, with 6 strata defined by case/control status, SES (high/low) and prevalence (high/low).

   The data set provided, **Hpylori_extract. dta,** is an extract from the full data, containing only children from low prevalence countries, with the following information: H. Pylori (HP) status of the child (**hpproband**), HP status of the mother (**hpmother**), SES category (**sescat**) and an indicator variable (**stage2**) for whether the mother's HP infection status is known.

   Use a two-stage analysis to estimate the effect of a HP-positive mother on the child's risk, adjusted for SES. (Note that since we are investigating only children from low-prevalence countries, the second-stage data is in 4 strata defined by case/control and high/low SES). Compare your results to those from the more detailed model in the Table below, where the full data set was analysed using a weighted regression (the coloumn "complete" refers to a simple logistic regression of the complete (i.e. stage 2) data, i.e. only children whose family members were tested.

	Complete data (four schools)	Weighted (correct)
Mother Infected	11.6 (2.0, 67.9)	12.7 (3.1, 51.6)
Absent	- -	3.6 (0.4, 31.4)
Father Infected	1.4 (0.2, 9.8)	1.7 (0.5, 6.4)
Absent	1.4 (0.2, 10.8)	1.3 (0.3, 6.7)
Sib(s) infected	8.1 (1.8, 37.3)	10.3 (2.8, 38.4)
No sibs	5.2 (0.7, 38.1)	5.9 (1.3, 27.6)
SES	2.6 (0.6, 12.1)	2.1 (0.8, 5.9)
High prev origin	6.7 (1.7, 25.7)	3.6 (0.9, 14.0)
Antibiotics	0.6 (0.2, 2.3)	0.5 (0.2, 1.6)

(Optional)

3.

i. Repeat the analysis in Question 1 using a two-stage case-control study with 777 randomly sampled men and 777 randomly sampled women selected at the second stage.

ii. Repeat the two-stage sampling experiment again but this time taking a second-stage sample of 1554 (i.e. 2 X 777) that is "balanced" across the eight strata defined by the **anychd, sex** and **prevhyp** (i.e. equal numbers in each stratum) using the following two approaches:

   a. Select one-eight of the 1554 (195) from each stratum (or all, if a stratum has fewer)
   b. Strategy a. will result in fewer than 1554. Select the "surplus" from the strata larger than 195 to yield a sample of size 1554 that is as "balanced" as possible

Re-run the weighted logistic regression for second-stage data for these two designs, using the appropriate weights. Compare your estimates to those in Question 1.

Sampling strategy b. is just one way of allocating "surplus" records to achieve a given total sample size. However, these could have been chosen from any stratum/strata, provided they are chosen randomly: explain how this would impact on the analysis and suggest why an investigator might wish to take an "unbalanced" sample.

Hints on next page….

Hints for Stata users

Using standard Stata commands

To select the second-stage samples, you can use random uniform numbers (*remember to use a seed*) with the command **runiform().** For example, to select a completely random sample of 777 controls, generate a random uniform number using **runiform()** , sort the data by the outcome and this random number, and then select the first 777 records within the control group using the index _n.

To selecting a balanced sample within strata, first create a stratum identifier (e.g. 1,2,3…. 8 in the example above) and get the sizes of the strata using the index _N. Then use a random uniform number, but this time sort the data by stratum and the uniform number and select the first n within each stratum using the index _n.

The weights are the stratum sizes dived by the numbers selected in the strata. Note that for strata with insufficient observations, you will need to set the weight to 1, as these strata will be sampled 100%.

The weighted logistic regression is run with the weights as a "pweights"
(remember to request the robust variance option using "vce(robust)").

Purpose-built package for two-stage analysis

The **meanscor** package, which you were requested to install before the course, has special commands (**meanscor**) for analyzing two-stage data. For example, with the "meanscor" command, the user specifies: the logistic model, first stage variables defining the sampling strata, and the second stage variable(s) in the model.

You can get more information by typing

```
help meanscor
```

or the PDF file on course web

Hints for R users

Using standard R commands

To select the second-stage samples, you can use random uniform numbers (*remember to use a seed*) generated by **runif()**, sort the data and select the number of rows needed. For example:to select a completely random sample of 777 controls, generate a random uniform number using **runif()**, sort the data by case/control status and this random number. Group the sample by case/control status generating a row ID for each subject within a group, and then select the number required using **mutate**. Alternatively, using **dplyr** library you can use **slice_sample** function to select rows from a data randomly and specify the argument **n** to indicate the number of rows to select and let the argument, **replace=FALSE**, to indicate sampling without replacement.

To select a balanced sample across strata, first create a stratum identifier using the **interaction** command and if **dplyr** library is used, you could use **group_by(strat) %>% slice_sample(195, replace=FALSE)**. To get the stratum sizes, you could use the **table** function or if **dplyr** library is used, **group_by(strat) %>% mutate(stratsize=n())**.

The weights are the stratum sizes dived by the numbers selected in the strata. Note that for strata with insufficient observations, you will need to set the weight to 1, as these strata will be sampled 100%.

Assuming the two-stage sample is saved as a dataframe named **two_stage**, the weighted logistic regression of the second-stage data can be fitted using the **glm** command, with the weights computed above assigned to the **weights** option. Robust standard errors can be obtained with the **sandwich** package, using the following commands:

```
cov <- vcovHC(fitted_glm, type = "HC0")
robust.se <- sqrt(diag(cov))
```

Purpose-built package for two-stage analysis:

The purpose-built package called **survey** implements an appropriate analysis: the two-stage design is specified using the function **twophase** and the mean score approach for binary outcome is implemented by the function **svyglm**.