Exercise 4.2: Optimal sampling (For Stata software)

For conducting these exercises in Stata, you will need to have installed the **optimal** functions as explained in the course web page (https://www.meb.ki.se/~biostat/beyond_classic_2022.htm) in the advice for Stata users prior to the course. These functions calculate the following optimal two-stage designs:
optfixn: to select an optimal second-stage sample of a fixed size
optprec: to select the optimal first and second stage samples that minmise cost for a required precision for one coefficient
optbud: to select the optimal first and second stage samples for a specifed budget that minimize the variance of a coefficient.

Run the help files for examples of how to use the commands

Q1.
The data set **wtpilot.dta** contains the binary outcome (operative mortality: **mort**) and a number of demographic and clinical covariates (age: **age**, sex: **sex**, weight: **weight**, surgery: **surg**, angina: **angina**, chf: **chf**, lvedp: **lvedp**) for a pilot sample of size 118 selected from the full CASS data that was analysed in Reilly AJE 1996. The original plan for the pilot data was to sample 10 observations from each of the twelve strata defined by the outcome, sex, and a 3-category weight variable, but as seen in the following table from the paper, one category had only 8 observations:

TABLE 5. Numbers of subjects from the Coronary Artery Surgery Study, stratified by operative mortality, sex, and weight

Sex and weight (kg)	Alive	Deceased
Males		
<60	160	8
60–70	1,083	33
≥70	5,418	103
Females		
<60	440	18
60–70	407	26
≥70	378	14
Total	7,886	202

i. Run the **optfixn** command with the **wtpilot** data and the population numbers in the 12 strata from **Table 5** above, to obtain an optimal second-stage sample of 1000 which minimizes the variance of left ventricular end diastolic blood pressure (LVEDP) in the following logistic model:

mort = α + **sex** + **weight** + **age** + **angina** + **chf** + **lvedp** + **surg**

(You will note some differences from Table 6 in the AJE paper, which used a different pilot sample).

ii. The full cohort data from which the pilot sample was drawn is in the dataset **chap8_CASS.dta**. Selected another balanced pilot sample of the same size as in (i), and compare the optimal design obtained using this pilot data to the design in (i). (NOTE: Your sample size may be less than 1000 in the analysis because some cohort members are missing LVEDP!)

iii. If the study in (i) or (ii) cost an average of €1 for each first stage and each second stage observation, then the total cost would be approximately €10,000. Use the **optbud** command to find the optimal design for this budget if a first-stage observation costs €5 and a second stage observation costs €50. What is the total study size, the total and stratum-specific second-stage sample size, the total cost of this study and the variance achieved for the coefficient of LVEDP.

iv. Use the **optprec** command to design a study that would achieve the variance you found in (iii), for minimum cost, assuming the same costs as before for a first and second-stage observation (€5 and €50 respectively). What do you notice?

v. Show that for the same budget as in (iv), if C1 = €5 and C2 = €10 the optimal design is a study approximately 50% larger than in (iv), which would sample at least twice as many controls in each stratum.

Q2.

The data set **ectopic.dta** consist of a total of 979 observations from a case-control study of ectopic pregnancy, with the following variables:

y	a binary indicator for case (i.e. ectopic pregnancy) or control
gonn	history of gonorrhea infection
contr	use of contraception
sexptn	multiple sex partners
chlam	positive for chlamydia antibody (***available for 327 individuals***)
z	a stratum variable constructed from gonn, contr and sexptn.

i. Assusing the 327 individuals with chlamydia results were sampled from the strata defined by **gonn**, **contr** and **sexptn** (i.e. the variable "**z**"), analyse all the data using weighted logistic regression or the **meanscor** command, to estimate the effect of exposure to chlamydia antibody on the risk of ectopic pregnancy, adjusted for the other variables.

ii. Verify the numbers and % in the final column of the following table (Table 8.2 in CES 8.2) using suitable cross-tabulation commands (or the **coding** command in Stata).

Stratum	Case/ Control	Contraceptive Use	Multiple Sexual Partners	N	n (%)
1	0	No	No	56	13 (23%)
2	0	No	Yes	138	41 (30%)
3	0	Yes	No	146	36 (25%)
4	0	Yes	Yes	375	102 (27%)
5	1	No	No	26	11 (42%)
6	1	No	Yes	189	97 (51%)
7	1	Yes	No	9	5 (56%)
8	1	Yes	Yes	40	22 (55%)

iii. Verify that the adjusted OR for chlamydia is approximately the same if we assume the sampling depended only on contraceptive use and sex partners.

The remaining sections of this exercise will use these two (contr, sexptn) as first stage variables as in CES 8.2.

iv. Using the available 327 complete observations as pilot data, verify that for a sample of this size (i.e. **optfixn** with n=327), the minimum achievable variance for the coefficient of chlamydia is .0674. How does this compare to the actual variance achieved by the investigators with the data collected in the study?

v. Use the **optprec** command to verify that if the first and second stage observations cost €5 and €50 respectively, the design that achieves the same variance as in (iv) for minimum cost is:

First stage sample size: 672
Sampling fractions in strata **Z**=1,2,…8 = (.56, 1, .08, .22, .97, .77, 1, 1)