**Stata Assignment 3**

**Sarah Pickersgill**

Use the Stata datasets vipcls and vipcls1 to create new analysis variables and a permanent dataset in Stata (vipcls2). Use this new dataset to answer the subsequent questions. Coding for questions 1 and 2 should be incorporated into the “analysis.do” file that you created for assignment 2. Please turn in this updated do file (pasted into your word document), as well as answers or output corresponding to those items that are bolded below:

1. In Assignment #2, you identified various possible outliers in the VIPCLS dataset. However, so that we all have the same recoding for the rest of the semester, we need to all make the same decisions regarding these variables. Using VIPCLS and/or VIPCLS1, create a new dataset VIPCLS2 which contains only the following changes (make any necessary modifications to your “analysis.do” file):

All -1 values should be set to missing

Set mother’s age (MOMAGE) of 15 to missing

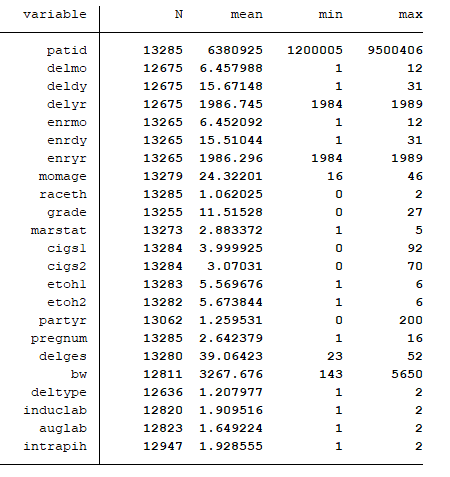
Set gestational age (DELGES) above 52 to missing

Set birth weight of babies (BW) below 300 to missing

Set birth weight of babies (BW) 6000 and above to missing

All other values for variables we will consider as correct

**Using summarize, produce output to summarize your new VIPCLS2 Stata Dataset** in order to double check that your dataset has the appropriate number of observations and values**.**

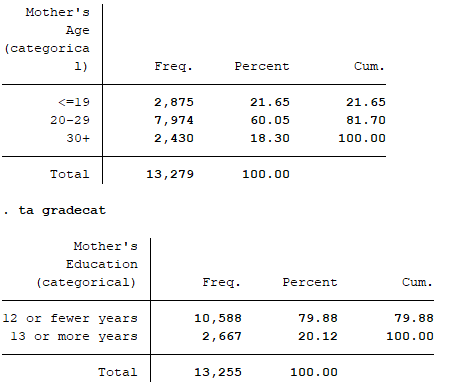


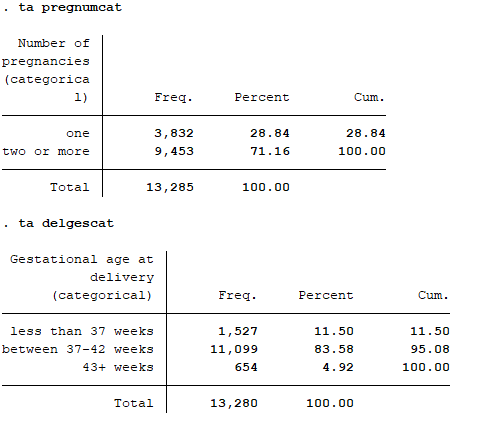
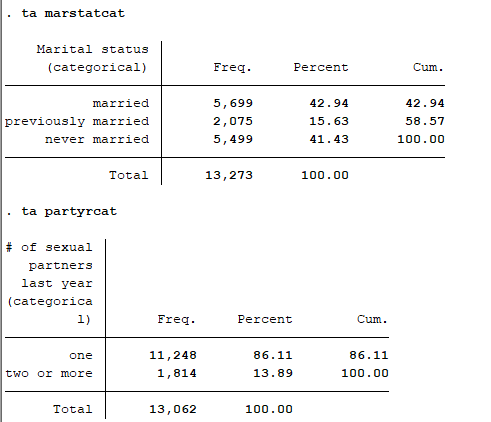
1. Working with the dataset VIPCLS2, define new categorical variables based upon the variables MOMAGE (<19, 20-29, 30+), GRADE (<12, 13+), MARSTAT (married, previously married, never married), PARTYR (<1, 2+), PREGNUM (1 vs 2+), DELGES (<37, 37-42, 43+), and BW (<1500, 1500-2499, 2500-3999, 4000+). In addition, define 2 new variables categorizing smoking during pregnancy (never smoked, 1st trimester only, 2nd trimester with or without smoking in the 1st trimester), and alcohol use during pregnancy (non-drinker, use in 1st trimester only, use in 2nd trimester with or without use in the 1st trimester). Finally, create a new variable for Study Clinic, which is based upon the PATID (Clinic 1 has PATID 1,000,000 to 1,999,999, Clinic 3 has PATID from 3,000,000 to 3,999,999, etc). Clinics (which are fabricated) are as follows: 1=Olympia, 3=Everett, 5=Seattle, 6=Bellingham, 7=Spokane, 8=Bellevue, 9=Tacoma

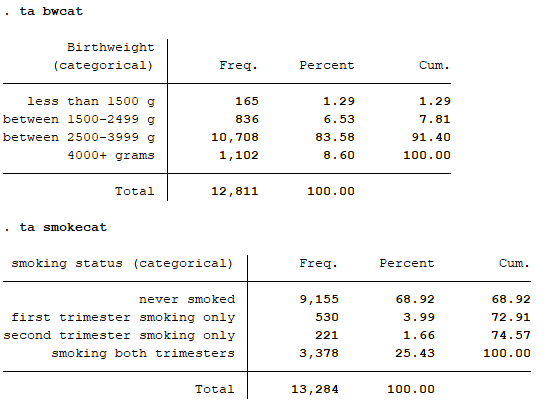
Add these variables to your permanent Stata dataset called VIPCLS2. Make sure you have labeled each new variable in the dataset.

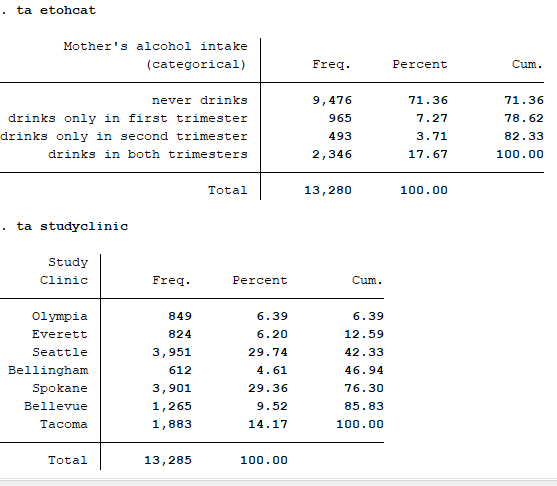
Create label definitions for your new variables.

**Use tabulate to obtain frequencies of the newly created recoded variables.** Again, make sure that all of the variables are labeled appropriately and that you have the appropriate missing values.

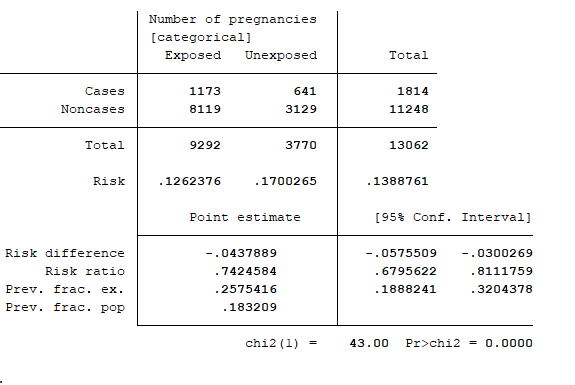








1. **Produce a 2x2 table and describe (in words and numbers) the relationship between having multiple sex partners in the last year and being in your first (vs. 2nd or later) pregnancy.**

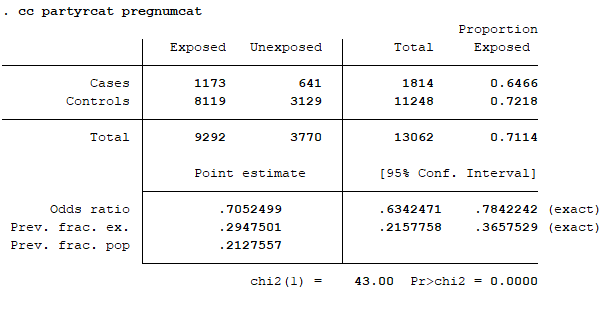


The probability of having multiple sexual partners among those with multiple pregnancies is 1173/9292 = (12.62%).

The probability of having multiple sexual partners among those in their first pregnancy is 641/3770 = (17.00%).

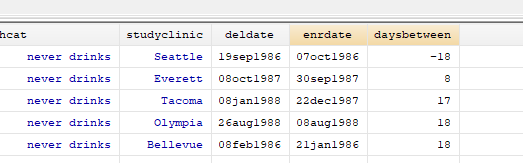
The ratio between those (the risk ratio) = 0.74 w/ 95% CI [0.67 , 0.81]. This mean that the likelihood of having more than one sexual partner in the last year is 26% less likely for those with multiple pregnancies compared to one pregnancy. The p-value for the risk ratio is < 0.0001.

Could alternatively do an odds ratio too:

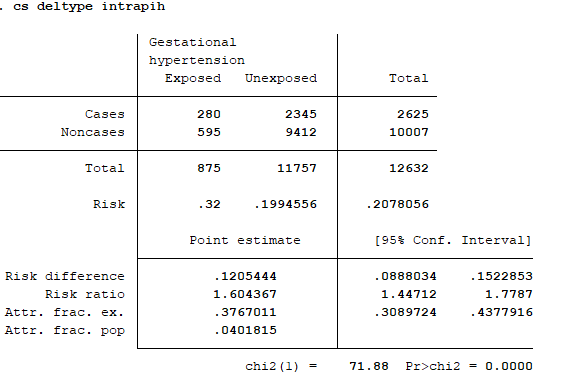


OR = 0.705 [0.63 , 0.78] with p-value < 0.0001. Meaning that women with multiple pregnancies (outcome) are less likely to have multiple sexual partners (exposure) than women in the first pregnancy. Odds are most useful for rare outcomes, or in a case-control study design.

1. Create two new variables for the enrollment date and the delivery date, and a new variable for the number of days between enrollment and delivery. Label these three variables and add them to your permanent datasets. **Determine if there are any delivery dates which precede enrollment. If so, at which clinic was the subject who had date problems enrolled? S**et any implausible number of days between enrollment and delivery to missing in your permanent Stata dataset.

There seems to be one value for which the delivery date precedes the enrollment date. This implausible value comes from Everett clinic. 

1. **Produce an appropriate 2x2 table (with labels), obtain a crude risk estimate and 95% confidence interval, and give a one sentence answer** to assess each of the following relationships. Think carefully about which factor is your exposure and which is your outcome, and also which are the appropriate reference groups:
   1. Is gestational hypertension associated with increased risk for delivery by c-section?

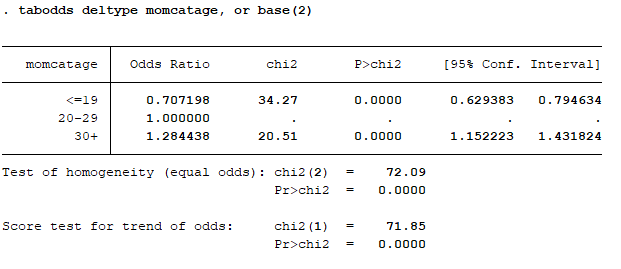


cases= c-section & exposure = hypertension

RR = 1.6 [1.4 , 1.8] with p-value < 0.0001

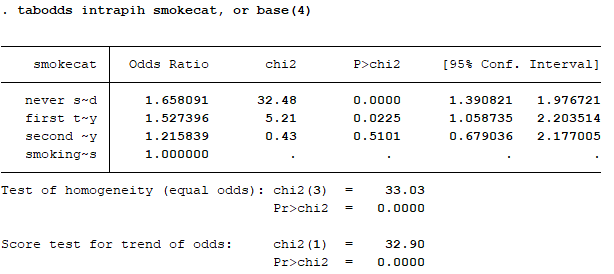
Meaning, there is a 1.6 times higher risk of c-section for mothers with hypertension compared to those who do not have hypertension.

* 1. Do older women (≥30) deliver by cesarean section more often than women in their 20’s?



OR = 1.28 [1.15 ,1.43] with p-value < 0.0001. Meaning, the odds of c-section for mothers age 30 and above is 1.28 higher compared to mothers in their 20’s.

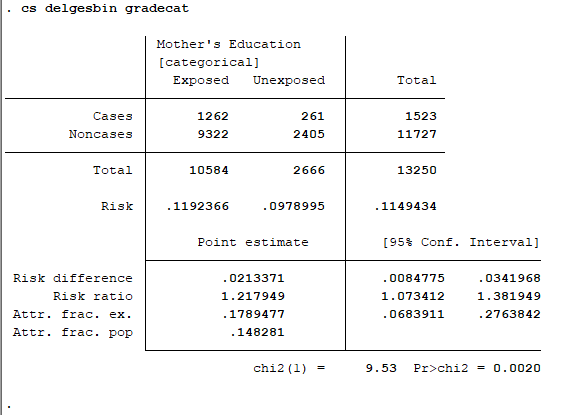
* 1. Is quitting smoking associated with a reduction in gestational hypertension? (Note: think carefully about the appropriate reference group for the exposure)



The reference group is those who smoked in both trimester 2 and 1 (represented above at smoking~s). I will compare this to those who only smoked in the first trimester (first t~y) as those would be the mothers who quit smoking.

OR = 1.527 [1.06 , 2.20] ] with p-value < 0.0225. Meaning, the odds of hypertension for mothers who quit smoking after the first trimester is 1.53 higher compared to mothers who smoked both trimesters. However, the p-values for this trend are relatively high.

* 1. Is having less education (≤12 years as in HWK #2) associated with having a preterm baby (create a dichotomous variable for preterm delivery (<37 weeks vs. 37+ weeks))?

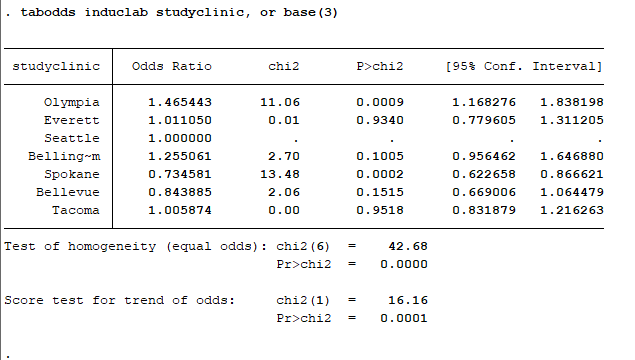


Cases = preterm delivery (<37 weeks) and exposure=less than 12 yrs education

RR = 1.12 [1.07 , 1.38] with p-value = 0.0020

Meaning, there is a 1.12 times higher risk of pre-term delivery for mothers with 12 years of less education compared to those who have over 12 year eduction.

* 1. Do Hispanic women enrolled in Spokane have their labor induced more often than Hispanic women enrolled in Seattle?



The reference group is Seattle. Cases=those who had induced labour.

OR = 0.734 [0.622 , 0.867 ] with p-value = 0.0002. Meaning, the odds of induced labor for mothers went to the clinic in Spokane is 0.734 as high as those mothers who went to a clinic in Seattle.