Assignment 4
Sampled Teenage Sexual Activity and Drug/Alcohol Use Data
Due November 10, 2015

Description of Data

Recall our in class analysis of the teen sex and drug/alcohol use dataset, and our discussion of missing data and concerns that both sexual activity and diary recording might be related to a teen's overall compliance. (That is, we were concerned that teens more likely to provide complete data would be less likely to engage in drinking or sexual activity). I mentioned that one way to assess that might be to sample from the data to obtain an equal number of entries for each respondent. The file teensex3.csv is data that has been processed from the original data. Among all respondents who reported at least 3 days of drug/alcohol use and sexual activity data, we sampled 3 records per person.

There are many ways one can analyze this data. This exercise involves doing several of them and trying to understand how they are similar and different, with regards to both the estimates and the inference. To do so, conduct the following analyses, with one sentence at the end of each analysis (1-4) summarizing your findings.

1. Do a random effects logistic regression model allowing for a subject-specific intercept. (In Stata, melogit can do this)
2. Find a marginal OR using GEE with independent and exchangeable correlations structures, with and without robust standard errors.
3. Provide a summary odds ratio and risk difference. Try using the cs or cc commands in STATA.
4. Use a paired t-test to test the difference in outcomes and interpret results. What is the parameter of interest implied by t-test? Is it the same or different than the OR provided by logistic regression?
5. Now that you have completed all 4 analyses, provide a summary table of your estimates and standard errors. Except for the ttest, provide your results in OR form. (What does the ttest provide?) Write a paragraph interpreting the differences and similarities among the results of the different analyses, including the assumptions of the techniques, the implied parameter of interest, the standard errors of the estimate of the parameters. What do we assume in sampling the data? What biases may still be present?
6. We have provided a table below that recaps the analysis of the full dataset presented in class. Compare your results to the results of that analysis. What do the differences suggest? What do we gain and what do we lose by sampling from our data?

For question 6: Summary of analysis of full data
(109 individuals who reported drug/alcohol use on the same day at least once)

	Estimate (OR/mean dif)	SE
ttest	-.118	.0246
melogit	1.474	.229
xtgee, cor(ind)	1.740	.202
xtgee , cor(ind), ro	1.740	.315
xtgee, cor(exch)	1.394	.170
xtgee, cor(exch), ro	1.394	.192