Assignment 4 Key
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 and meqrlogit can do this)

```
. meqrlogit sx24hrs drgalcoh || eid:, or


Mixed-effects logistic regression               Number of obs      =       294
Group variable: eid                             Number of groups   =        98

                                                Obs per group: min =         3
                                                               avg =       3.0
                                                               max =         3

Integration points =   7                        Wald chi2(1)       =      1.41
Log likelihood = -168.84216                     Prob > chi2        =    0.2347

------------------------------------------------------------------------------
     sx24hrs | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    drgalcoh |   1.661568   .7099111     1.19   0.235     .7191823    3.838815
       _cons |   .2758797   .0907653    -3.91   0.000     .1447677    .5257361
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
eid: Identity                |
                  var(_cons) |   4.236699   1.615505      2.006568     8.94543
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =    36.62 Prob>=chibar2 = 0.0000
```

The random effects model estimate that for each teenage, the odds of having sex on a day when they drank or did drugs are 1.66 times the odds of having sex on a day they did note drink or have drugs. This is a subject-specific effect.

2. Find a marginal OR using GEE with independent and exchangeable correlations structures, with and without robust standard errors.

   The estimates can be seen in the table below. The commands to use are:

   ```
   xtgee sx24hrs drgalcoh, family(binomial) i(eid) corr(ind) eform
   xtgee sx24hrs drgalcoh, family(binomial) i(eid) corr(ind) eform ro
   xtgee sx24hrs drgalcoh, family(binomial) i(eid) corr(exch) eform
   xtgee sx24hrs drgalcoh, family(binomial) i(eid) corr(exch) eform ro
   ```

   The marginal models estimate the population odds ratio for having sex on a day when drugs and alcohol have been consumed vs a day when drugs and alcohol have not been consumed.

   We use both an independent correlation structure, which assumes all observations are independent, and an exchangeable correlation structure, which assumes a correlation structure for each teen. (Teens are independent from each other.) The estimates of the odds ratios are similar. In each case, the robust standard errors are larger than the model-based standard errors. However, the difference between the model-based and robust standard errors is smaller when we use the exchangeable correlation structure, which indicates that model is a better fit to the data.

3. Provide a summary odds ratio and risk difference. Try using the cs or cc commands in STATA.

   ```
   . cc sx24hrs drgalcoh
                                                            Proportion
                    |   Exposed   Unexposed  |      Total     Exposed
   -----------------+------------------------+------------------------
              Cases |        35          64  |         99       0.3535
           Controls |        56         139  |        195       0.2872
   -----------------+------------------------+------------------------
              Total |        91         203  |        294       0.3095
                    |                        |
                    |      Point estimate    |    [95% Conf. Interval]
                    |------------------------+------------------------
         Odds ratio |         1.357422       |   .7814252    2.342142 (exact)
    Attr. frac. ex. |         .2633094       |   -.279713    .5730404 (exact)
    Attr. frac. pop |         .0930892       |
                    +-------------------------------------------------
                                  chi2(1) =     1.35  Pr>chi2 = 0.2448
   ```

   ```
   . cs sx24hrs drgalcoh

                    | drgalcoh               |
                    |   Exposed   Unexposed  |      Total
   -----------------+------------------------+------------
              Cases |        35          64  |         99
           Noncases |        56         139  |        195
   -----------------+------------------------+------------
   ```

```
          Total |         91          203  |          294
                |                          |
           Risk |   .3846154     .3152709  |     .3367347
                |                          |
                |       Point estimate     |    [95% Conf. Interval]
                |------------------------+------------------------
Risk difference |         .0693444         |    -.0493002      .187989
     Risk ratio |         1.219952         |     .8773962     1.696249
Attr. frac. ex. |         .1802956         |     -.139736      .410464
Attr. frac. pop |         .0637409         |
                +-------------------------------------------------
                              chi2(1) =      1.35  Pr>chi2 = 0.2448
```

As expected, the summary odds ratio is the same as given by the logistic regression assuming independence of observations. The risk difference, a different parameter, shows an effect in the same direction.

4. Use a 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?

```
. ttest(sx24hrs), by(drgalcoh)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     203    .3152709    .0326908    .4657723    .2508119    .3797299
       1 |      91    .3846154    .0512821    .4891996    .2827346    .4864961
---------+--------------------------------------------------------------------
combined |     294    .3367347    .0276092    .4733991    .2823972    .3910722
---------+--------------------------------------------------------------------
    diff |           -.0693444    .0596861               -.186814    .0481251
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.1618
Ho: diff = 0                                     degrees of freedom =      292

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1231         Pr(|T| > |t|) = 0.2463          Pr(T > t) = 0.8769
```

The test again finds the risk difference, and shows it is not signficant.

5. Now that you have completed all analyses in Questions 1-4, 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?

Table: Analysis of sample of 3 observations per id, sampled w/o replacement from 98 individuals who had at least 3 observations

	OR	*Mean Difference*	*SE*
ttest	*--*	*-.069*	*.060*
meqrlogit	*1.66*	*--*	*.71*
xtgee, cor(ind)	*1.36*	*--*	*.36*
xtgee , cor(ind), ro	*1.36*	*--*	*.43*
xtgee, cor(exch)	*1.38*	*--*	*.36*
xtgee, cor(exch), ro	*1.38*	*--*	*.39*

The results from this analysis follow a pattern we would expect with a logistic regression. The individual effect estimated by the random effects model was larger than the population average effect estimated by GEE in the marginal models. The standard errors estimated robustly are larger than model based standard errors, as we expect, and it appears that the GEE model with exchangeable correlation is a better fit to the data than the model assuming independence of observations.

We sampled the data to address he unbalanced data structure we discussed in class. We were concerned that more compliant individuals – those who reported every day for the full month – were less likely to have sex than those who reported for fewer days. This meant that observations those who had less sex were overrepresented in the dataset. (We first noticed this because of the unusual pattern of results shown in the table for question 6: the marginal model with an independent correlation structure gave a larger estimate than the random effects (individual-level model). Additionally, the use of an exchangeable correlation structure reduced the estimate substantially)

In sampling 3 observations per individual, we attempt to provide a more balanced data structure and eliminate the over-representation of the more compliant individuals. However, in doing this, we discard the individuals with fewer than 3 observations, which means we are excluding the least compliant individuals from the analysis. Additionally, we are implicitly assuming that for those individuals who had fewer than 30 observations, those observations that they do provide are representative of their full month. If an individual's behavior influenced his or her decision to report, sampling 3 observations per person will not correct for that bias.

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

As discussed above, by sampling the data, we appear to do a pretty good job of correcting the problems introduced by the unbalanced nature of the original dataset. The estimates produced in the analysis of the sampled data are more likely to answer our research question about the relationship between substance use and sexual activity among teenagers in this study. However, by sampling from the data, we do lose power as reduce our sample size.