Assignment 4

John Semerdjian

November 10, 2015

Sampled Teenage Sexual Activity and Drug/Alcohol Use Data

1. Do a random effects logistic regression model allowing for a subject-specific intercept.

```
library(geepack)
library(lme4)
library(lmtest)
library(sandwich)

df = read.csv("../data/teensex3.csv")

# default cunstructured cor for mixed effects in glmer
# glmer does not return robust SE (no package that I'm aware of can do this in R)
mixed_model = glmer(sx24hrs ~ drgalcoh + (1|eid), data = df, family="binomial")
mixed_or = exp(fixef(mixed_model))
mixed_se = sqrt(diag(vcov(mixed_model)))
cbind("OR"=mixed_or, "Naive SE"=mixed_se)
```

```
##                     OR  Naive SE
## (Intercept) 0.2811666 0.3167558
## drgalcoh    1.6344398 0.4151650
```

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

```
# independent
gee_ind_model = geeglm(sx24hrs ~ drgalcoh, id=eid, data=df,
                       family="binomial", corstr="independence")
gee_ind_or = exp(coef(gee_ind_model))
gee_ind_naive_se = sqrt(diag(gee_ind_model$geese$vbeta.naiv))
gee_ind_robust_se = sqrt(diag(gee_ind_model$geese$vbeta))
cbind("OR"=gee_ind_or, "Robust SE"=gee_ind_robust_se, "Naive SE"=gee_ind_naive_se)
```

```
##                     OR Robust SE  Naive SE
## (Intercept) 0.4604317 0.1803540 0.1510604
## drgalcoh    1.3574219 0.3176786 0.2631498
```

```
# exchangeable
gee_exc_model = geeglm(sx24hrs ~ drgalcoh, id=eid, data=df,
                       family="binomial", corstr="exchangeable")
gee_exc_or = exp(coef(gee_exc_model))
gee_exc_naive_se = sqrt(diag(gee_exc_model$geese$vbeta.naiv))
gee_exc_robust_se = sqrt(diag(gee_exc_model$geese$vbeta))
cbind("OR"=gee_exc_or, "Robust SE"=gee_exc_robust_se, "Naive SE"=gee_exc_naive_se)
```

```
##                     OR Robust SE  Naive SE
## (Intercept) 0.4575868 0.1832561 0.1864278
## drgalcoh    1.3840153 0.2802246 0.2567811
```

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

The SE of the summary OR was calculated using the following formula:

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

```
# results below calculated using data from a 2x2 table:
# with(df, table(drgalcoh, sx24hrs))
odds_ratio = (139/64)/(56/35)
odds_ratio_se = sqrt(1/139 + 1/64 + 1/56 + 1/35)
risk_difference = 139/(139+56)-64/(64+35)

rbind("Summary OR"=odds_ratio,
      "Summary OR SE"=odds_ratio_se,
      "Summary Risk Diff"=risk_difference)
```

```
##                         [,1]
## Summary OR        1.35742188
## Summary OR SE     0.26314980
## Summary Risk Diff 0.06635587
```

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?

A t-test is testing the difference in means between two samples. The test returns a t statistic and a 95% confidence interval around the mean difference may be calculated. We're looking at the risk difference and not the odds ratio. The t-test does not return an odds ratio.

The summary OR calculated in the previous question would be similar to the OR calculated from a logistic regression, however.

I calculated the SE of a two sample t-test using the following formula:

$$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

```
ttest = t.test(sx24hrs ~ drgalcoh, data=df, var.equal=FALSE)
ttest_mean_diff = ttest$estimate[1] - ttest$estimate[2]

yes_drgalcoh = df[df$drgalcoh == 1, "sx24hrs"]
no_drgalcoh = df[df$drgalcoh == 0, "sx24hrs"]

ttest_SE = sqrt(var(yes_drgalcoh)/length(yes_drgalcoh) +
                var(no_drgalcoh)/length(no_drgalcoh))

cbind(list("t-test Mean Diff"=ttest_mean_diff,
           "t-test SE"=ttest_SE))
```

```
##                   [,1]
## t-test Mean Diff -0.06934445
## t-test SE         0.0608156
```

5. Now that you have completed all 4 analyses, provide a summary table of your estimates and standard errors. Except for the t-test, provide your results in OR form. (What does the t-test 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 the sampling data? What biases may still be present?

As mentioned in the previous question, the t-test tests the difference between the means in two samples.

GEE is used when we are interested in the population average effect. We make few assumptions when fitting a GEE model. An exchangeable structure assumes that all measurements for an individual have the same ρ. Mixed effects models are used to understand effects at the individual level. We assume that we know the structure of the data generating process; we also assume that we know the distribution of the random effect, subject id (`eid`), in our case.

The OR and SE estimates from the mixed effects model were the largest of all the analyses. Mixed effects models in R assume an unstructured correlation by default. The OR estimates from the exchangeable GEE models (naive and robust) were slightly larger than the independent GEE OR estimates. The SE estimates from independent GEE were slightly larger than their respective counterparts in the exchangeable GEE models. The summary OR and SE estimates were identical to indepedent GEE. As expected, the robust SE were larger than the naive SE between the GEE models.

In our sample, we sample 3 observations per subject. The assumption we make is that the samples were drawn randomly, which is likely wrong since the mean number of observations per group is unequal. There could be a difference between subjects that report data for all days vs. subjects that report data for only a few days, which may bias our results.

```
result_est = rbind("Summary"=odds_ratio,
  "t-test"=ttest_mean_diff, "Mixed Effects"=mixed_or[2],
  "GEE cor(Ind)"=gee_ind_or[2], "GEE cor(Ind) robust"=gee_ind_or[2],
  "GEE cor(Exch)"=gee_exc_or[2], "GEE cor(Exch) robust"=gee_exc_or[2]
  )

result_se = rbind(odds_ratio_se,
  ttest_SE, mixed_se[2],
  gee_ind_naive_se[2], gee_ind_robust_se[2],
  gee_exc_naive_se[2], gee_exc_robust_se[2]
  )

results = data.frame(result_est, result_se)
colnames(results) = c("Estimate (OR/Mean Diff)", "SE")
results
```

```
##                       Estimate (OR/Mean Diff)        SE
## Summary                            1.35742188 0.2631498
## t-test                            -0.06934445 0.0608156
```

```
## Mixed Effects                  1.63443982 0.4151650
## GEE cor(Ind)                   1.35742188 0.2631498
## GEE cor(Ind) robust            1.35742188 0.3176786
## GEE cor(Exch)                  1.38401529 0.2567811
## GEE cor(Exch) robust           1.38401529 0.2802246
```

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 our data?

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

	Estimate (OR/Mean Diff)	SE
t-test	-0.118	.025
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

The mixed effects model from the sampled data had a larger OR and SE compared to the mixed effects model built on the full data set. This difference is largely due to the missing data in the original data set. We removed more observations from subjects that did not take drugs or drink than those that did.

The OR for the indepedent GEE models (naive and robust) on the sampled dataset was much smaller, while the SE's were only slightly smaller than the full data set SE's. The OR from the exchangeable GEE model from the sampled data set were slightly smaller than the full data set's OR, while both naive and robust SE's were larger. The similarity in the OR results between the GEE models from the sampled data show that there was an large impact from creating a subset of equal observations between subjects. The correlation structure did not reveal a large difference between the OR's.

The difference in the means and their SE (from t-test) in the sampled data were much smaller than the full data set results.

In summary, the differences are largely due to our assumptions of randomness in our sample. In our attempt to control for the missing data in original data set by looking 3 complete observations per subject, we also introduce bias.