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(gee)
library(lme4)
# library(multcomp) # for STATA lincom functionality
library(lmtest)   # for robust errors
library(sandwich) # for robust errors

df = read.csv("../data/teensex3.csv")
df$today = as.Date(df$today, "%d-%b-%y")

# default covar structure is unstructured
mixed_model = glmer(sx24hrs ~ drgalcoh + (1|eid), data = df, family="binomial")
mixed_or = exp(fixef(mixed_model))
mixed_se = sqrt(diag(vcov(mixed_model)))
# mixed effects in R do not return robust SE
cbind("OR"=mixed_or, "Naive SE"=mixed_se)
```

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

```
# unit of analysis is teen, bootstrap doesn't account for this
# confint(mixed_model, method="bootstrap")
# exp(glht(mixed_model, linfct = "drgalcoh = 0")$coef)
```

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
```

```r
# 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.

   What is risk difference? Odds ratios are not good effect measures when the outcome of interest is common. Odds are close to risks when the outcome is rare, but exaggerated (odds >> risk) when the outcome is common. Thus, odds ratio tends to give false impression of large effect when the outcome is common.

   Risk difference (attributable risk) - difference in proportions

```r
# calculate or by hand or for all models?
# with(df, table(drgalcoh, sx24hrs))
# Odds ratio:                (139/64)/(56/35)
# Risk difference:           139/(139+56) - 64/(64+56)
# Absolute Risk Reduction:  64/(139+64)  - 35/(56+35)
# library(epiR)
# epi.2by2(with(df, table(drgalcoh, sx24hrs)), conf.level = 0.95)
```

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?

```r
glm_model = glm(sx24hrs ~ drgalcoh, data = df, family="binomial")

glm_or = exp(coef(glm_model))
glm_naive_se = sqrt(diag(vcov(glm_model)))
glm_robust_se = sqrt(diag(vcovHC(glm_model, "HC1")))
cbind("OR"=glm_or, "Naive SE"=glm_naive_se, "Robust SE"=glm_robust_se)
```

```
##                    OR  Naive SE Robust SE
## (Intercept) 0.4604317 0.1510604 0.1515768
## drgalcoh    1.3574219 0.2631498 0.2640495
```

```r
# t-test of what outcomes?
```

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?

```r
cbind("Mixed Effects"=mixed_or, "GEE Indep."=gee_ind_or, "GEE Exch."=gee_exc_or)
```

```
##             Mixed Effects GEE Indep. GEE Exch.
## (Intercept)     0.2811666  0.4604317 0.4575868
## drgalcoh        1.6344398  1.3574219 1.3840153
```

```r
# analyses?
# 1. mixed
# 2. gee ind
# 3. gee ech
# 4. t-test
```

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?