

BIOS 525 Final Exam Data Analysis Project Report

Jiawei Meng

Question 1: Diarrheal Prevalence

The dataset we studied contains 5464 study participants from 152 schools. There are around 36 participants in each school, except school id 136 has 8 participants. There are no missing records, so we include all the participants in our study, and they were analyzed descriptively the result is show in Table 1 below. The mean student's age at the time of survey was 10.6 years old. About half of the participants are involved in each treatment group. We found that less students had diarrhea (9.2%) over the last three days. And 90.8% participants didn't have diarrhea over the last three days.

Table 1. Summary statistics for all covariates at individual and/or school levels

	No treat (N=2738)	Have treat (N=2726)	Overall (n=5464)
Student's Sex			
Boy	1395 (50.9%)	1382 (50.7%)	2777 (50.8%)
Girl	1343 (49.1%)	1344 (49.3%)	2687 (49.2%)
Student's Age at the time of survey			
Mean (SD)	10.6 (1.80)	10.8 (1.80)	10.7 (1.77)
Median [Min, Max]	11.0 [7.00, 15.0]	11.0 [7.00, 15.0]	11.0 [7.00, 15.0]
Whether the child had diarrhea			
No	2265 (82.7%)	2696 (98.9%)	4961 (90.8%)
Yes	473 (17.3%)	30 (1.1%)	503 (9.2%)

We examined whether the programs for safe water, adequate sanitation, and improved hygiene have effect on diarrhea prevalence between schools. Since the students in one school will share a same condition, so observations within a school between different students are likely to be more similar than those between different schools. But we are concerned about the effect of the hygiene program variable on the diarrhea outcome on average in all school, so we fit a generalized linear model using generalized estimating equations. This approach can obtain school average effect accounting for the fact that in one school the observations has some correlation

We decided to fit the generalized estimation equation model with exchangeable correlation, because we think that in one school, the correlation between observations for two students is equal to any two students. Since the hygiene program variable is the primary interest, we first fit the GEE model only include the treat variable, the result shows that the p-value for treat variable is less than 0.05 so we think there's some statistically significant association between hygiene programs and diarrhea prevalence, and the estimator for treat is less than 0 so the hygiene programs has negative effect on the diarrhea prevalence, which means using the program will

decrease the diarrhea prevalence. Then we checked whether the association was modified by age and sex by adding two interaction terms treat multiply age and treat multiply sex to the primary model. The result shows both age and sex interaction term are significant, which means the association between hygiene programs and diarrhea prevalence were modified by age and sex. But when we add both 'treat*age' and 'treat*sex' interaction term, the treat program, participants' age and sex become not significant on diarrhea prevalence, which is not consistent with our primary model so it may be not a good idea to add both two interaction, after checking one by one, we found that when we just contain the treat and sex interaction term, the covariates are all significant, although the covariate for sex is marginally significant (larger than but very close to 0.05), we decided to keep treat, sex, and their interaction in our model, then we got the final model:

$$\text{logit } P(Y_{ij} = 1) = \beta_0 + \beta_1 \text{treat}_i + \beta_2 \text{sex}_{ij} + \beta_5 \text{treat}_i * \text{sex}_{ij}$$

We can conclude from the final model that hygiene program is benefit for the school to prevent diarrhea and the effect of treat on diarrhea prevalence significantly modified by sex.

The results of the estimates of coefficients are showed in table 2. Based on our final model, assuming an exchangeable correlation for the diarrhea outcome observed within the same school, we estimated a population-average OR in girls for hygiene program $\exp(-2.505) = 0.082$ (95%CI: 0.052, 0.129) (Table 3) using a robust standard error. And the population-average OR in boys for hygiene program $\exp((-2.505) + (-1.021)) = 0.029$ (95%CI: 0.009, 0.098) using a robust standard error. Therefore, our result suggests an association between hygiene program and diarrhea prevalence, the using of hygiene program will decrease the diarrhea prevalence in the school and the association will modified by sex.

Table 2. Coefficients of generalized estimation equation model

covariate	Estimate	Naïve S.E.	Naïve. z	Robust S.E.	Robust z	P-value
(Intercept)	-1.664	0.091	-18.22	0.098	-16.94	<0.01***
Treat	-2.505	0.289	-8.67	0.232	-10.81	<0.01***
Sex	0.191	0.100	1.91	0.099	1.94	0.0522*
Treat:Sex	-1.021	0.415	-2.46	0.383	-2.66	<0.01***

Table 3. Odds Ratio for covariate treat

covariate	Odds Ratio Estimate	Lower CI	Upper CI
Treat (Girl)	0.082	0.052	0.129
Treat (Boy)	0.029	0.009	0.098

Question 2: Chlamydia Incidence

The dataset we studied contains 1310 study participants from 131 counties. There are 10 reporting time each year for every county. There are no missing records, so we include all the observations in our study, and they were analyzed descriptively the result is show in Table 1 below. The mean Median household income from Census 2000 is 48400. The mean percent black population from Census 2000 is 14.2%, the mean percent Hispanic population from Census 2000 is 14.9%. The number of reporting year for each county is 10 from 2003 to 2012.

Table 1. Summary statistics for all covariates at county levels

	Overall (N=131)
Median household income (in \$1,000) from Census 2000	
Mean (SD)	48.4 (11.0)
Median [Min, Max]	46.0 [25.0, 81.0]
Percent black population from Census 2000	
Mean (SD)	14.2 (13.0)
Median [Min, Max]	10.0 [0, 64.0]
Percent Hispanic population from Census 2000	
Mean (SD)	14.9 (14.9)
Median [Min, Max]	10.0 [1.00, 88.0]
Number of Reporting for each county	
Mean (SD)	10.0 (0)
Median [Min, Max]	10.0 [10.0, 10.0]

We examined whether population characteristics are associated with chlamydia incidence rates. The chlamydia cases from the same county were clustered together. Since each county level will have their variation, we can use random intercept model to find out how much of an effect county's population has on chlamydia incidence.

We decided to fit the random intercept model and first we created a variable 'rate' to indicate the chlamydia incidence rates, which equals to the cases divided by the at-risk population size. Then the outcome data 'rate' has two levels which is the county level i: 1 to 13 and year level j: 1 to 10. We include all the covariates to the primary model and check the significance of the covariates according to the p-value. Except covariate percent Hispanic population has p-value larger than 0.05, other covariates are all significant. So we don't include the percent Hispanic population covariates and the final model is as below:

$$Y_{ij} = \beta_{0,i} + \beta_{1i}year_{ij} + \beta_{2i}HHIncome2000_i + \beta_{3i}PBlack2000_i + \epsilon_{ij}, \beta_{0,i} = \mu + \theta_i, \\ \theta_i \sim N(0, \tau^2), \epsilon_{ij} \sim N(0, \sigma^2),$$

We can conclude from the final model that the median household income and the percent of black population are significantly associated with chlamydia incidence rates. The results of the estimates of coefficients are showed in table 2. Based on our final model, after controlling for county-specific baseline chlamydia incidence rates, we estimated a one-year increase will result in 1.96e-04 increase in chlamydia incidence rates. A unit increase in average median household

income will result in -9.04e-05 increase in chlamydia incidence rates. A unit increase in average percent black population will result in 9.62e-05 increase in chlamydia incidence rates. Therefore, our result suggests an association between population characteristics household income and black population percent and chlamydia incidence rates, the increase in household income and the decrease in black population percentage will decrease the chlamydia incidence rates

Table 2. Coefficients of random intercept model

Covariates	Estimate	Std. Error	P-value
Intercept	6.19e-03	5.63e-04	<0.01***
Year	1.96e-04	6.71e-06	<0.01***
HHIncome2000	-9.04e-05	1.04e-05	<0.01***
PBlack2000	9.62e-05	8.75e-06	<0.01***

Question 3: Power Calculations

Assume that the intervention does not change over time, the Regression model $Y_{ij} = \theta_i + \mu_i + \beta_1 * \text{intervention}_{ij} + \epsilon_{ij}$, $\theta_i \sim N(0, \tau^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$ Code the intervention term: 0 = control, 1 = intervention, for each subject, there are 3 observations.

The minimal detectable effect size with 80% power and a type I error rate of 0.05 based on the above model is 1.4.

If the investigator interested in estimating how the duration of the intervention may impact its effect, we will create new variable 'time' to indicate the duration of intervention for that observation. Code the intervention term: 0 = control, 1 = intervention, for each subject, there are 3 observations. Code the time term: 0 = intervention duration is 0-year, 1 = intervention duration is 1 year, 2 = intervention duration is 2 year, 3 = intervention duration is 3 year. Our model is $Y_{ij} = \theta_i + \mu_i + \beta_1 * \text{intervention}_{ij} + \beta_2 * \text{time}_{ij} + \epsilon_{ij}$, $\theta_i \sim N(0, \tau^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$

The minimal detectable effect size with 80% power and a type I error rate of 0.05 based on the above model is 2.06.

(The procedure in in the code attached.)

Appendix

Jiawei Meng

Question 1: Diarrheal Prevalence

```
### import data
diarrhea <- read.csv ("/Users/mengjiawei/Desktop/2020fall/525/Final/diarrhea.csv",header=T)

# check observations
dim(diarrhea)
```

```
## [1] 5464    5
```

```
# check missing data
summary(diarrhea)
```

```
##      School      Sex      Age      treat
## Min.   : 1.00   Min.   :0.0000   Min.   : 7.00   Min.   :0.0000
## 1st Qu.: 38.00   1st Qu.:0.0000   1st Qu.: 9.00   1st Qu.:0.0000
## Median : 76.00   Median :0.0000   Median :11.00   Median :0.0000
## Mean   : 76.31   Mean    :0.4918   Mean    :10.73   Mean    :0.4989
## 3rd Qu.:115.00   3rd Qu.:1.0000   3rd Qu.:12.00   3rd Qu.:1.0000
## Max.   :152.00   Max.    :1.0000   Max.    :15.00   Max.    :1.0000
##      Z
## Min.   :0.00000
## 1st Qu.:0.00000
## Median :0.00000
## Mean    :0.09206
## 3rd Qu.:0.00000
## Max.    :1.00000
```

```
# the number of school in the data
length(unique(diarrhea$School))
```

```
## [1] 152
```

```
# the number of observations in each school
table(diarrhea$School)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
## 38 36 37 36 27 37 33 39 38 27 33 39 36 38 38 39 40 40 35 40
## 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
## 36 40 34 39 37 40 39 38 36 32 34 33 34 39 25 40 40 35 32 39
## 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
```

```
## 35 38 38 37 34 32 34 34 35 37 36 39 37 39 38 36 38 37 38 38
## 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
## 36 39 38 37 38 40 35 40 35 35 35 31 40 26 35 33 34 37 36 39
## 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
## 33 37 29 38 37 38 38 38 36 33 39 34 40 39 35 18 23 29 32 39
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
## 34 39 37 26 37 36 38 38 39 39 38 35 40 38 34 38 38 35 37 40
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
## 40 37 40 40 39 35 22 38 37 40 38 37 38 37 38 8 39 40 37 12
## 141 142 143 144 145 146 147 148 149 150 151 152
## 40 30 38 40 35 37 39 40 39 40 37 39
```

```
### descriptive statistics
diarrhea$Sex<-factor(diarrhea$Sex,levels=c(0,1),
                    labels = c("boy","girl"))
diarrhea$treat<-factor(diarrhea$treat,levels=c(0,1),
                      labels = c("no","yes"))
diarrhea$Z<-factor(diarrhea$Z,levels=c(0,1),
                   labels = c("no","yes"))

library(table1)
```

```
##
## Attaching package: 'table1'
```

```
## The following objects are masked from 'package:base':
##
## units, units<-
```

```
label(diarrhea$Sex)<-"Student's sex"
label(diarrhea$School)<-"School ID"
label(diarrhea$Age)<-"Student's age at the time of survey"
label(diarrhea$treat)<-"the presence of a school-level program to improve sanitation and hygiene"
label(diarrhea$Z)<-"whether the child had diarrhea over the last three days"
table1(~ Sex + Age + Z |treat, data = diarrhea)
```

```
## [1] "<table class=\\"Rtable1\\">\n<thead>\n<tr>\n<th class='rowlabel firstrow lastrow'></th>\n<th class='>
```

```
# instal required package
library(gee)
library(geepack)
# load data and factor variable
diarrhea <- read.csv ("/Users/mengjiawei/Desktop/2020fall/525/Final/diarrhea.csv",header=T)

diarrhea$Sex<-factor(diarrhea$Sex,levels=c(0,1),
                    labels = c("boy","girl"))
diarrhea$treat<-factor(diarrhea$treat,levels=c(0,1),
                      labels = c("no","yes"))

# check interaction
#treat
fittrt = gee(Z~treat ,family = binomial(link = "logit"),data=diarrhea,
            id=School, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)    treatyes
##   -1.566235    -2.932092
```

```
2* pnorm(abs(coef(summary(fitttrt))[,5]), lower.tail = FALSE)
```

```
## (Intercept)    treatyes
## 4.125626e-71 3.862183e-43
```

```
#age
```

```
fitage = gee(Z~treat + Age,family = binomial(link = "logit"),data=diarrhea,
            id=School, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)    treatyes      Age
## -1.02123021 -2.92449286 -0.05147299
```

```
2* pnorm(abs(coef(summary(fitage))[,5]), lower.tail = FALSE)
```

```
## (Intercept)    treatyes      Age
## 2.237879e-03 2.915154e-41 7.816947e-03
```

```
fitagei = gee(Z~treat * Age,family = binomial(link = "logit"),data=diarrhea,
            id=School, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)    treatyes      Age treatyes:Age
## -1.27237145 1.04742955 -0.02768844 -0.39028006
```

```
2* pnorm(abs(coef(summary(fitagei))[,5]), lower.tail = FALSE)
```

```
## (Intercept)    treatyes      Age treatyes:Age
## 7.666812e-05 3.608877e-01 7.659542e-02 9.522839e-04
```

```
#sex
```

```
fitsex = gee(Z~ treat + Sex,family = binomial(link = "logit"),data=diarrhea,
            id=School, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)    treatyes    Sexgirl
## -1.6339397 -2.9331701 0.1349931
```

```
2* pnorm(abs(coef(summary(fitsex))[,5]), lower.tail = FALSE)
```

```
## (Intercept)      treatyes      Sexgirl
## 2.981478e-65 3.084568e-43 1.970888e-01
```

```
fitsexi = gee(Z~treat * Sex,family = binomial(link = "logit"),data=diarrhea,
             id=School, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
```

```
## (Intercept)      treatyes      Sexgirl treatyes:Sexgirl
## -1.6700093      -2.5014433      0.2045884      -1.0325978
```

```
2* pnorm(abs(coef(summary(fitsexi))[,5]), lower.tail = FALSE)
```

```
## (Intercept)      treatyes      Sexgirl treatyes:Sexgirl
## 2.322373e-64      3.080778e-27      5.222765e-02      7.709005e-03
```

```
#two inter
fittwo = gee(Z~ treat * Sex + treat* Age,family = binomial(link = "logit"),
             data=diarrhea, id=School, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
```

```
## (Intercept)      treatyes      Sexgirl      Age
## -1.38221033      1.51192790      0.20354180      -0.02706698
## treatyes:Sexgirl      treatyes:Age
## -1.04558663      -0.39304982
```

```
summary(fittwo)
```

```
##
## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link:                      Logit
## Variance to Mean Relation: Binomial
## Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = Z ~ treat * Sex + treat * Age, id = School, data = diarrhea,
##      family = binomial(link = "logit"), corstr = "exchangeable")
##
## Summary of Residuals:
##      Min      1Q      Median      3Q      Max
## -0.213708243 -0.163365312 -0.017203874 -0.004882499 0.995117501
```



```

##
##
## Coefficients:
##           Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept) -1.1573407 0.30981911 -3.735537  0.25910069 -4.466760
## treatyes    1.4640173 1.20814006  1.211794  1.12948349  1.296183
## Sexgirl     0.1878651 0.09718309  1.933105  0.09789646  1.919018
## Age        -0.0476058 0.02824849 -1.685251  0.02723911 -1.747700
## treatyes:Sexgirl -1.0246072 0.39836996 -2.571999  0.37819639 -2.709194
## treatyes:Age   -0.3875886 0.12610183 -3.073616  0.11752044 -3.298052
##
## Estimated Scale Parameter: 0.9468744
## Number of Iterations: 3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 1.00000000 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [2,] 0.03460657 1.00000000 0.03460657 0.03460657 0.03460657 0.03460657
## [3,] 0.03460657 0.03460657 1.00000000 0.03460657 0.03460657 0.03460657
## [4,] 0.03460657 0.03460657 0.03460657 1.00000000 0.03460657 0.03460657
## [5,] 0.03460657 0.03460657 0.03460657 0.03460657 1.00000000 0.03460657
## [6,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 1.00000000
## [7,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [8,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [9,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [10,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [11,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [12,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [13,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [14,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [15,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [16,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [17,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [18,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [19,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [20,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [21,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [22,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [23,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [24,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [25,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [26,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [27,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [28,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [29,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [30,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [31,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [32,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [33,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [34,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [35,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [36,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [37,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657
## [38,] 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657 0.03460657

```

[illegible]

[illegible]

[illegible]

[illegible]

```
## [9,] 0.03460657 0.03460657 0.03460657 0.03460657
## [10,] 0.03460657 0.03460657 0.03460657 0.03460657
## [11,] 0.03460657 0.03460657 0.03460657 0.03460657
## [12,] 0.03460657 0.03460657 0.03460657 0.03460657
## [13,] 0.03460657 0.03460657 0.03460657 0.03460657
## [14,] 0.03460657 0.03460657 0.03460657 0.03460657
## [15,] 0.03460657 0.03460657 0.03460657 0.03460657
## [16,] 0.03460657 0.03460657 0.03460657 0.03460657
## [17,] 0.03460657 0.03460657 0.03460657 0.03460657
## [18,] 0.03460657 0.03460657 0.03460657 0.03460657
## [19,] 0.03460657 0.03460657 0.03460657 0.03460657
## [20,] 0.03460657 0.03460657 0.03460657 0.03460657
## [21,] 0.03460657 0.03460657 0.03460657 0.03460657
## [22,] 0.03460657 0.03460657 0.03460657 0.03460657
## [23,] 0.03460657 0.03460657 0.03460657 0.03460657
## [24,] 0.03460657 0.03460657 0.03460657 0.03460657
## [25,] 0.03460657 0.03460657 0.03460657 0.03460657
## [26,] 0.03460657 0.03460657 0.03460657 0.03460657
## [27,] 0.03460657 0.03460657 0.03460657 0.03460657
## [28,] 0.03460657 0.03460657 0.03460657 0.03460657
## [29,] 0.03460657 0.03460657 0.03460657 0.03460657
## [30,] 0.03460657 0.03460657 0.03460657 0.03460657
## [31,] 0.03460657 0.03460657 0.03460657 0.03460657
## [32,] 0.03460657 0.03460657 0.03460657 0.03460657
## [33,] 0.03460657 0.03460657 0.03460657 0.03460657
## [34,] 0.03460657 0.03460657 0.03460657 0.03460657
## [35,] 0.03460657 0.03460657 0.03460657 0.03460657
## [36,] 0.03460657 0.03460657 0.03460657 0.03460657
## [37,] 1.00000000 0.03460657 0.03460657 0.03460657
## [38,] 0.03460657 1.00000000 0.03460657 0.03460657
## [39,] 0.03460657 0.03460657 1.00000000 0.03460657
## [40,] 0.03460657 0.03460657 0.03460657 1.00000000
```

```
2* pnorm(abs(coef(summary(fittwo))[,5]), lower.tail = FALSE)
```

```
##      (Intercept)      treatyes      Sexgirl      Age
## 7.941303e-06    1.949125e-01    5.498204e-02    8.051598e-02
## treatyes:Sexgirl    treatyes:Age
## 6.744697e-03    9.735799e-04
```

```
#final
fit1 = gee(Z ~ treat + Sex + treat * Sex, family = binomial(link = "logit"),
           data=diarrhea, id=School, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
```

```
##      (Intercept)      treatyes      Sexgirl treatyes:Sexgirl
## -1.6700093    -2.5014433    0.2045884    -1.0325978
```

```
summary(fit1)
```

```
##
## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link:                      Logit
## Variance to Mean Relation: Binomial
## Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = Z ~ treat + Sex + treat * Sex, id = School, data = diarrhea,
##      family = binomial(link = "logit"), corstr = "exchangeable")
##
## Summary of Residuals:
##      Min      1Q      Median      3Q      Max
## -0.18654711 -0.15926135 -0.01523867 -0.00670503  0.99329497
##
##
## Coefficients:
##              Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept)   -1.6637343 0.09133035 -18.216663  0.09821945 -16.938949
## treatyes      -2.5048287 0.28899311  -8.667434  0.23171025 -10.810176
## Sexgirl        0.1911301 0.09994387   1.912374  0.09845709   1.941253
## treatyes:Sexgirl -1.0207368 0.41539658  -2.457259  0.38307952  -2.664556
##
## Estimated Scale Parameter:  0.9997308
## Number of Iterations:  2
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 1.00000000 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [2,] 0.03196691 1.00000000 0.03196691 0.03196691 0.03196691 0.03196691
## [3,] 0.03196691 0.03196691 1.00000000 0.03196691 0.03196691 0.03196691
## [4,] 0.03196691 0.03196691 0.03196691 1.00000000 0.03196691 0.03196691
## [5,] 0.03196691 0.03196691 0.03196691 0.03196691 1.00000000 0.03196691
## [6,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 1.00000000
## [7,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [8,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [9,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [10,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [11,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [12,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [13,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [14,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [15,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [16,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [17,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [18,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [19,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [20,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [21,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
```

[illegible]

[illegible]

[illegible]

[illegible]

```
## [33,] 0.03196691 0.03196691 1.00000000 0.03196691 0.03196691 0.03196691
## [34,] 0.03196691 0.03196691 0.03196691 1.00000000 0.03196691 0.03196691
## [35,] 0.03196691 0.03196691 0.03196691 0.03196691 1.00000000 0.03196691
## [36,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 1.00000000
## [37,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [38,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [39,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
## [40,] 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691 0.03196691
##      [,37]      [,38]      [,39]      [,40]
## [1,] 0.03196691 0.03196691 0.03196691 0.03196691
## [2,] 0.03196691 0.03196691 0.03196691 0.03196691
## [3,] 0.03196691 0.03196691 0.03196691 0.03196691
## [4,] 0.03196691 0.03196691 0.03196691 0.03196691
## [5,] 0.03196691 0.03196691 0.03196691 0.03196691
## [6,] 0.03196691 0.03196691 0.03196691 0.03196691
## [7,] 0.03196691 0.03196691 0.03196691 0.03196691
## [8,] 0.03196691 0.03196691 0.03196691 0.03196691
## [9,] 0.03196691 0.03196691 0.03196691 0.03196691
## [10,] 0.03196691 0.03196691 0.03196691 0.03196691
## [11,] 0.03196691 0.03196691 0.03196691 0.03196691
## [12,] 0.03196691 0.03196691 0.03196691 0.03196691
## [13,] 0.03196691 0.03196691 0.03196691 0.03196691
## [14,] 0.03196691 0.03196691 0.03196691 0.03196691
## [15,] 0.03196691 0.03196691 0.03196691 0.03196691
## [16,] 0.03196691 0.03196691 0.03196691 0.03196691
## [17,] 0.03196691 0.03196691 0.03196691 0.03196691
## [18,] 0.03196691 0.03196691 0.03196691 0.03196691
## [19,] 0.03196691 0.03196691 0.03196691 0.03196691
## [20,] 0.03196691 0.03196691 0.03196691 0.03196691
## [21,] 0.03196691 0.03196691 0.03196691 0.03196691
## [22,] 0.03196691 0.03196691 0.03196691 0.03196691
## [23,] 0.03196691 0.03196691 0.03196691 0.03196691
## [24,] 0.03196691 0.03196691 0.03196691 0.03196691
## [25,] 0.03196691 0.03196691 0.03196691 0.03196691
## [26,] 0.03196691 0.03196691 0.03196691 0.03196691
## [27,] 0.03196691 0.03196691 0.03196691 0.03196691
## [28,] 0.03196691 0.03196691 0.03196691 0.03196691
## [29,] 0.03196691 0.03196691 0.03196691 0.03196691
## [30,] 0.03196691 0.03196691 0.03196691 0.03196691
## [31,] 0.03196691 0.03196691 0.03196691 0.03196691
## [32,] 0.03196691 0.03196691 0.03196691 0.03196691
## [33,] 0.03196691 0.03196691 0.03196691 0.03196691
## [34,] 0.03196691 0.03196691 0.03196691 0.03196691
## [35,] 0.03196691 0.03196691 0.03196691 0.03196691
## [36,] 0.03196691 0.03196691 0.03196691 0.03196691
## [37,] 1.00000000 0.03196691 0.03196691 0.03196691
## [38,] 0.03196691 1.00000000 0.03196691 0.03196691
## [39,] 0.03196691 0.03196691 1.00000000 0.03196691
## [40,] 0.03196691 0.03196691 0.03196691 1.00000000
```

```
2* pnorm(abs(coef(summary(fit1))[,5]), lower.tail = FALSE)
```

```
##      (Intercept)      treatyes      Sexgirl treatyes:Sexgirl
##      2.322373e-64      3.080778e-27      5.222765e-02      7.709005e-03
```

```
## the estimate and se from above output
est = c(-1.664,-2.505, 0.191,-1.021)
se = c(0.0982,0.2317, 0.0985,0.3831)

## treatgirl
est_tg = est[2]+est[4]
se_tg = sqrt((se[2])^2 + 2*se[2]*se[4]*(1) +(se[4])^2)

# OR for treat in boy
ORb = exp(est[2])
lCIb = exp(est[2]-1.96*se[2])
uCIb = exp(est[2]+1.96*se[2])
b<- cbind(ORb,lCIb, uCIb)
# OR for treat in girl
ORg = exp(est_tg)
lCIg = exp(est_tg-1.96*se_tg)
uCIg = exp(est_tg+1.96*se_tg)
g<- cbind(ORg,lCIg, uCIg)
#OR table
resu<-rbind(b,g)
resu
```

```
##          ORb          lCIb          uCIb
## [1,] 0.08167560 0.051863916 0.12862321
## [2,] 0.02942237 0.008817578 0.09817615
```

```
rownames(resu) = c("treatboy","treatgirl")
colnames(resu) = c("est","lower","upper")
round(resu,3)
```

```
##          est lower upper
## treatboy  0.082 0.052 0.129
## treatgirl 0.029 0.009 0.098
```

Question 2: Chlamydia Incidence

```
cdc <- get(load('/Users/mengjiawei/Desktop/2020fall/525/Final/CDC.RData'))
dim(cdc)
```

```
## [1] 1310    9
```

```
### descriptive statistics
library(table1)
label(cdc$PBlack2000)<-"Percent black population from Census 2000"
label(cdc$PHisp2000)<-"Percent Hispanic population from Census 2000"
label(cdc$Cases)<-"Reported cases of chlamydia"
label(cdc$HHIncome2000)<-"Median household income (in $1,000) from Census 2000"
label(cdc$Population)<-"At-risk population size"

library(tidyverse)
```

```

## -- Attaching packages ----- tidyverse

## v ggplot2 3.3.2      v purrr  0.3.4
## v tibble  3.0.3      v dplyr  1.0.2
## v tidyr   1.1.2      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0

## -- Conflicts ----- tidyverse
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

#check duplicate and missing
nrow(cdc)

## [1] 1310

cdc = cdc %>%
  distinct() %>%
  filter(!is.na(Cases))
nrow(cdc)

## [1] 1310

#find number of county enrolled
length(unique(cdc$FIPS)) #131

## [1] 131

#check year
unique(cdc$Year)

## [1] 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012

# descriptive statistics
# table the income, black population, hispanic population in county level
table1(~ HHIncome2000 + PBlack2000 + PHisp2000, data = cdc
       %>% arrange(FIPS,Year) %>% filter(!duplicated(FIPS)))

## [1] "<table class='Rtable1'>\n<thead>\n<tr>\n<th class='rowlabel firstrow lastrow'></th>\n<th class="
cdcf1 = cdc %>%
  group_by(Year) %>%
  summarise(nid = n())

## `summarise()` ungrouping output (override with `.groups` argument)

table1(~nid,data=cdcf1)

## [1] "<table class='Rtable1'>\n<thead>\n<tr>\n<th class='rowlabel firstrow lastrow'></th>\n<th class="

```

```
#check the number of visit for each county
cdcf2 = cdc %>%
  group_by(FIPS) %>%
  summarise(nvisit=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
table1(~ nvisit , data=cdcf2)
```

```
## [1] "<table class=\"Rtable1\">\n<thead>\n<tr>\n<th class='rowlabel firstrow lastrow'></th>\n<th clas
```

```
#fit model
#center year
cdc$Year <- cdc$Year - min(cdc$Year)

#generate rate
cdc$Rate <- cdc$Cases/cdc$Population

library(lmerTest)
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##   expand, pack, unpack
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##   lmer
```

```
## The following object is masked from 'package:stats':
##
##   step
```

```
#fit model
fit.ran = lmer(Rate~(1|FIPS) + Year + HHIncome2000 + PBlack2000 + PHisp2000,
               data=cdc)
summary(fit.ran)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Rate ~ (1 | FIPS) + Year + HHIncome2000 + PBlack2000 + PHisp2000
## Data: cdc
##
## REML criterion at convergence: -14772.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.3399  -0.3795  -0.0016   0.3853  12.1418
##
## Random effects:
## Groups Name Variance Std.Dev.
## FIPS (Intercept) 1.496e-06 0.0012232
## Residual 4.869e-07 0.0006978
## Number of obs: 1310, groups: FIPS, 131
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)
## (Intercept) 6.186e-03 6.657e-04 1.275e+02 9.293 5.27e-16 ***
## Year 1.964e-04 6.712e-06 1.178e+03 29.253 < 2e-16 ***
## HHIncome2000 -8.884e-05 1.114e-05 1.270e+02 -7.971 7.82e-13 ***
## PBlack2000 9.730e-05 9.207e-06 1.270e+02 10.568 < 2e-16 ***
## PHisp2000 3.038e-06 7.972e-06 1.270e+02 0.381 0.704
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
## (Intr) Year HHI200 PB2000
## Year -0.045
## HHIncom2000 -0.948 0.000
## PBlack2000 -0.554 0.000 0.376
## PHisp2000 -0.530 0.000 0.361 0.301
```

```
fit.ran2 = lmer(Rate~(1|FIPS) + Year + HHIncome2000 + PBlack2000,
               data=cdc)
summary(fit.ran2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Rate ~ (1 | FIPS) + Year + HHIncome2000 + PBlack2000
## Data: cdc
##
## REML criterion at convergence: -14794.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.3399  -0.3795  -0.0011   0.3838  12.1446
##
## Random effects:
## Groups Name Variance Std.Dev.
## FIPS (Intercept) 1.486e-06 0.0012190
## Residual 4.869e-07 0.0006978
## Number of obs: 1310, groups: FIPS, 131
```



```
##
## Fixed effects:
##           Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  6.320e-03  5.627e-04  1.287e+02  11.231 < 2e-16 ***
## Year         1.964e-04  6.712e-06  1.178e+03  29.253 < 2e-16 ***
## HHIncome2000 -9.037e-05  1.036e-05  1.280e+02  -8.725 1.22e-14 ***
## PBlack2000    9.624e-05  8.751e-06  1.280e+02  10.997 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) Year   HHI200
## Year         -0.054
## HHIncom2000 -0.957  0.000
## PBlack2000  -0.488  0.000  0.300
```

Question 3: Power Calculations

a) Regression model $y_{ij} = \mu + \theta_i + \beta_1 \times intervention_{ij} + \epsilon_{ij}$. Code the intervention term: 0 = control, 1 = intervention, for each subject in each group, there are 3 observations.

b)

```
## Define model parameters:
m = 250
n = 3
rho = 0.3
tau2=12^2*rho
sigma2=12^2-tau2
alpha = 0.05

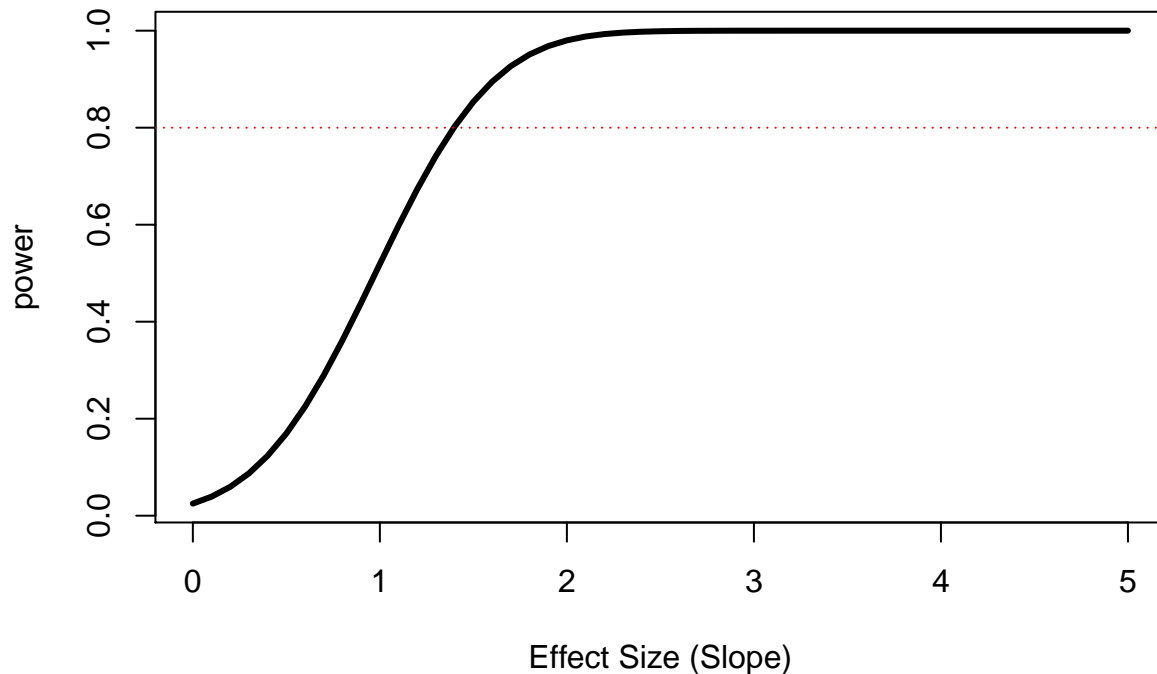
## A sequence of effect sizes (slope) to calculate power for
effs = seq(0, 5, by = 0.1)

## Set up design matrix
X = cbind(rep(1,m*3*n), c(rep(1, m*n), rep(c(0,1,1), m),
                           rep(c(0,0,1), m)))

## Set up covariance matrix
R0 = matrix(tau2,n,n)
diag(R0) <- sigma2 + tau2
V <- kronecker(diag(1,m*3), R0)

##Covariance of beta_hat and standard errors
VCOV = solve(t(X)%*%solve(V)%*%X)
SE = sqrt(diag(VCOV))

## From Slide 10
power = pnorm(qnorm(alpha/2) + abs(effs)/SE[2])
plot(power~effs, type = "l", lwd = 3, xlab = "Effect Size (Slope)")
abline(h = 0.8, col = 2, lty= 3)
```



```
MDE = (qnorm(0.8)-qnorm(alpha/2))*SE[2]
MDE
```

```
## [1] 1.395524
```

The minimum detectable effect size between intervention and control with 80% power and a two-sided type I error rate of 0.05 is 1.4.

- c) If the investigator interested in estimating how the duration of the intervention may impact its effect, we will create new variable 'time' to indicate the duration of intervention for that observation. Code the intervention term: 0 = control, 1 = intervention, for each subject, there are 3 observations. Code the time term: 0 = intervention duration is 0 year, 1 = intervention duration is 1 year, 2 = intervention duration is 2 year, 3 = intervention duration is 3 year. Regression model $y_{ij} = \mu + \theta_i + \beta_1 \times intervention_{ij} + \beta_2 \times time_{ij} + \epsilon_{ij}$.

```
## Define model parameters:
m = 250
n = 3
rho = 0.3
tau2=12^2*rho
sigma2=12^2-tau2
alpha = 0.05

## A sequence of effect sizes (slope) to calculate power for
effs = seq(0, 5, by = 0.1)

## Set up design matrix
X = cbind(rep(1,m*3*n), c(rep(1,m*n), rep(c(0,1,1), m),
                           rep(c(0,0,1), m)), c(rep(c(1,2,3), m), rep(c(0,1,2), m),
```

```

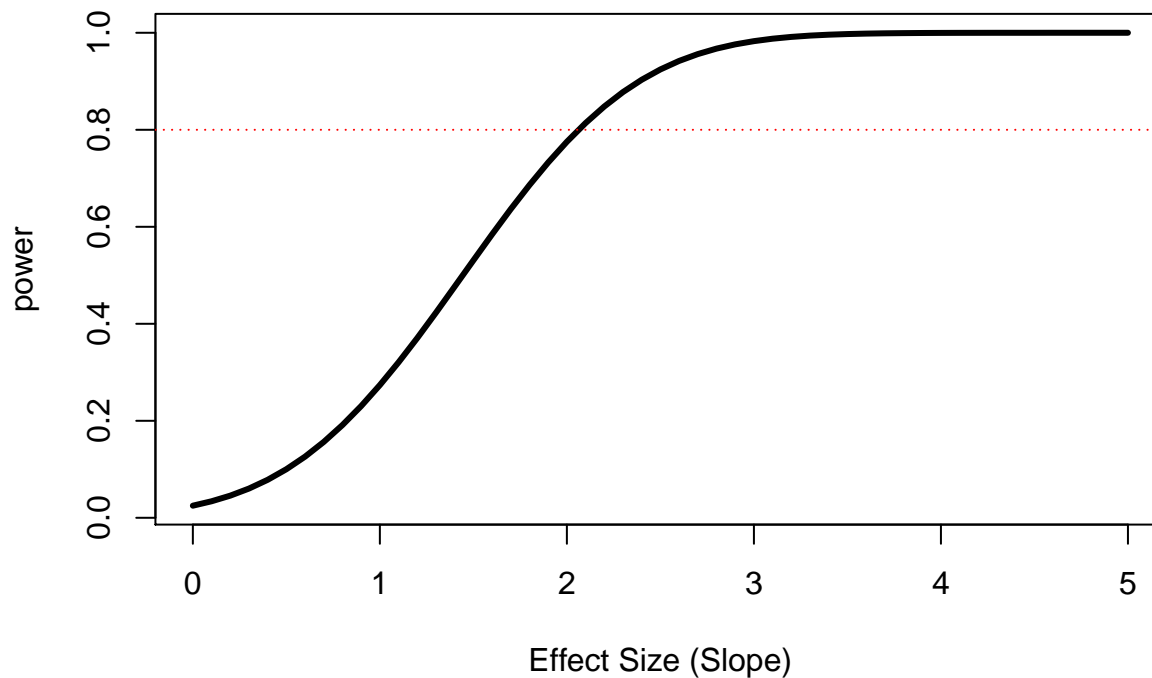
                                rep(c(0,0,1), m)))

## Set up covariance matrix
R0 = matrix(tau2,n,n)
diag(R0) <- sigma2 + tau2
V <- kronecker(diag(1,m*3), R0)

## Covariance of beta_hat and standard errors
VCOV = solve (t(X)%*%solve(V)%*%X)
SE = sqrt (diag (VCOV))

## From Slide 10
power = pnorm ( qnorm (alpha/2) + abs(efss)/SE[2])
plot (power~efss, type = "l", lwd = 3, xlab = "Effect Size (Slope)")
abline(h = 0.8, col = 2, lty= 3)

```



```

MDE = (qnorm(0.8)-qnorm(alpha/2))*SE[2]
MDE

```

```
## [1] 2.062695
```

The minimum detectable effect size between treatment and placebo with 80% power and a two-sided type I error rate of 0.05 is 2.06.