Homework 4

2023-02-14

Question 1: This exercise involves the analysis of Rao's bone growth data of 20 boys. The data are available on the class web page in the data files, rao.xls (Excel), rao.csv (CSV), and rao.dta (Stata). The variables are boy (boy ID number), age (1=8.0 years, 2=8.5 years, 3=9.0 years, 4=9.5 years), ht (ramus height in mm).

```
rao <- read.csv("C:/Users/sdm98/Desktop/BIOSTAT718/rao.csv", header = TRUE, sep = ',')

rao$age_yr <- NA
rao$age_yr[rao$age == 1] <- 8
rao$age_yr[rao$age == 2] <- 8.5
rao$age_yr[rao$age == 3] <- 9
rao$age_yr[rao$age == 4] <- 8.5
```

Part a) Estimate the rate of bone growth per year using a marginal model. Give a 95% confidence interval for the rate. Do three separate analyses, with (1) independence, (2) exchangeable correlation, and (3) a different working correlation of your choice (justify your choice).

```
#load gee library
library(geepack)
library(gee)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#run independent model
indep <- gee(ht ~ age_yr, id = boy, corstr = 'independence', data = rao)
```

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

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

```
## (Intercept)      age_yr
##     34.3625      1.8500
```

```
summary(indep)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Independent
##
## Call:
## gee(formula = ht ~ age_yr, id = boy, data = rao, corstr = "independence")
##
## Summary of Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0875 -2.0875 -0.0875  1.9875  5.9125
##
##
## Coefficients:
##             Estimate Naive S.E.  Naive z Robust S.E.  Robust z
## (Intercept)  34.3625  7.4000183 4.643570    2.432943 14.123842
## age_yr        1.8500  0.8698383 2.126832    0.285263  6.485243
##
## Estimated Scale Parameter:  7.566186
## Number of Iterations:  1
##
## Working Correlation
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1
```

```
#get confidence intervals
#age coefficient +/- 1.96*std error
lb <- indep$coefficients[2] - 1.96*0.285
ub <- indep$coefficients[2] + 1.96*0.285
paste("Estimated bone growth per year is", round(indep$coefficients[2],3), 'mm.')
```

```
## [1] "Estimated bone growth per year is 1.85 mm."
```

```
paste("Independence 95% CI = (",round(lb,3),',',round(ub,3),')')
```

```
## [1] "Independence 95% CI = ( 1.291 , 2.409 )"
```

```
#run exchangeable model
exch <- gee(ht ~ age_yr, id = boy, corstr = 'exchangeable', data = rao)
```

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

## running glm to get initial regression estimate

## (Intercept)      age_yr
##     34.3625      1.8500
```

```
summary(exch)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = ht ~ age_yr, id = boy, data = rao, corstr = "exchangeable")
##
## Summary of Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0875 -2.0875 -0.0875  1.9875  5.9125
##
##
## Coefficients:
##             Estimate Naive S.E.  Naive z Robust S.E.  Robust z
## (Intercept)  34.3625   3.617366 9.499316    2.432943 14.123842
## age_yr        1.8500   0.420468 4.399859    0.285263  6.485243
##
## Estimated Scale Parameter:  7.566186
## Number of Iterations:  1
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.7663375 0.7663375 0.7663375
## [2,] 0.7663375 1.0000000 0.7663375 0.7663375
## [3,] 0.7663375 0.7663375 1.0000000 0.7663375
## [4,] 0.7663375 0.7663375 0.7663375 1.0000000
```

```
#get confidence intervals
#age coefficient +/- 1.96*std error
lb_exch <- exch$coefficients[2] - 1.96*0.285
ub_exch <- exch$coefficients[2] + 1.96*0.285
paste("Estimated bone growth per year is", round(exch$coefficients[2],3), 'mm.')
```

```
## [1] "Estimated bone growth per year is 1.85 mm."
```

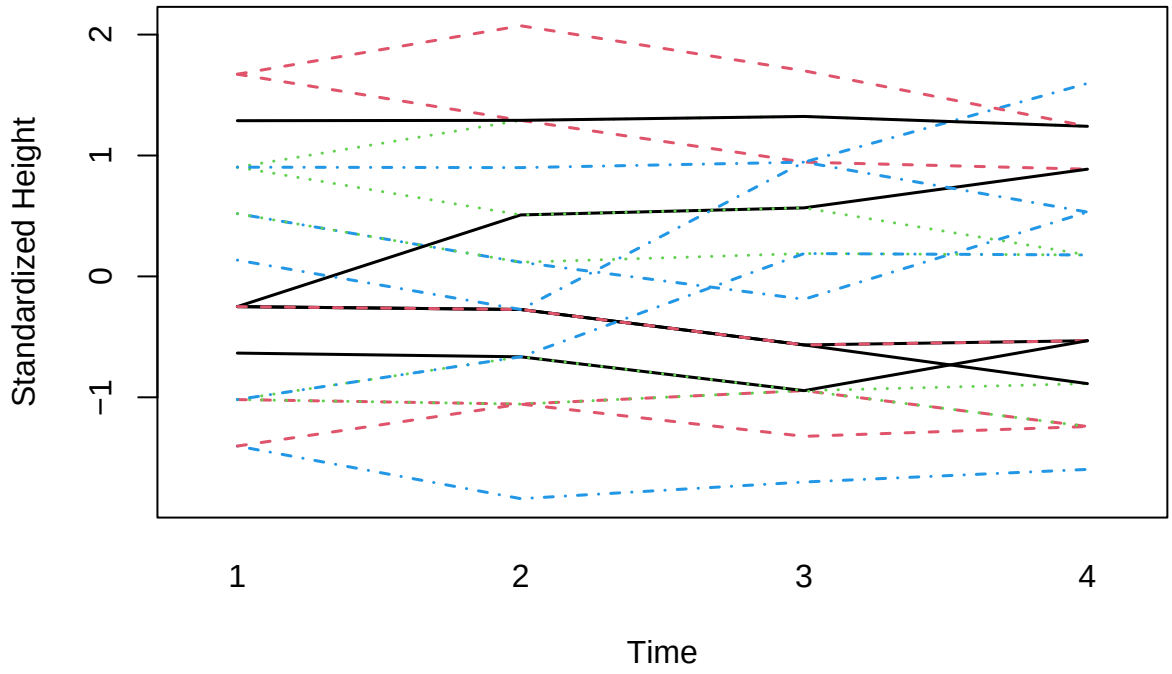
```
paste("Exchangeable 95% CI = (",round(lb_exch,3),',',round(ub_exch,3),')')
```

```
## [1] "Exchangeable 95% CI = ( 1.291 , 2.409 )"
```

```
#need to determine correlation structure to choose third model

#first, look at the change in height over time for each boy
mnht <- with(rao, tapply(ht,age,mean))
rao$mnht <- mnht[match(rao$age,names(mnht))]
sdht <- with(rao, tapply(ht, age, sd))
rao$sdht <- sdht[match(rao$age,names(sdht))]
rao$stdht <- (rao$ht - rao$mnht)/rao$sdht

attach(rao)
```

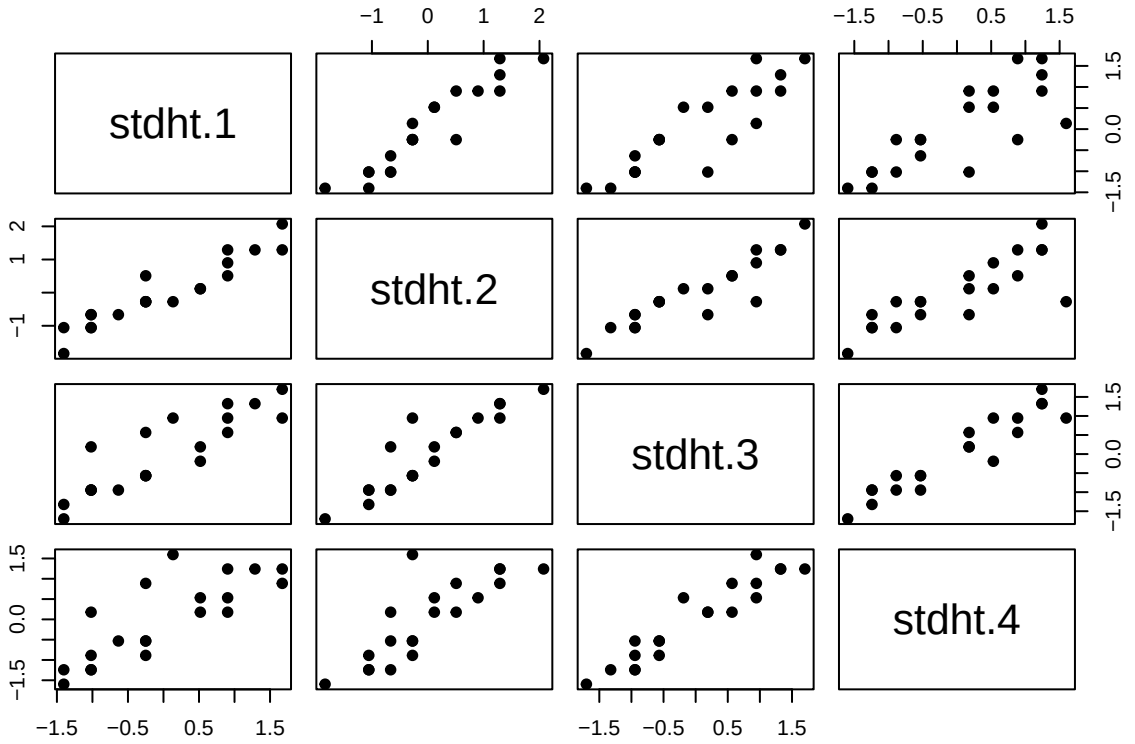
```
## The following objects are masked _by_ .GlobalEnv:
##
##     mnht, sdht
```

```
interaction.plot(age,boy,stdht,legend = F, col = 1:4,lty=1:4, xlab = 'Time',ylab= 'Standardized Height'
```




```
#next, examine residuals

#fit independent model to get residuals
mod <- lm(ht~age, data = rao)
rao$res <- mod$residuals

#plot scatter plots of residuals by time points
```

```
rao.sort <- rao[order(rao$boy),][,c(1,2,7)]
wide <- reshape(rao.sort,timevar = "age",idvar = c("boy"),direction="wide")

pairs(wide[,2:5],pch=19)
```




```
#get correlation table
round(cor(wide[,2:5]),4)
```

```
##         stdht.1 stdht.2 stdht.3 stdht.4
## stdht.1  1.0000  0.9410  0.8757  0.8075
## stdht.2  0.9410  1.0000  0.9181  0.8251
## stdht.3  0.8757  0.9181  1.0000  0.9456
## stdht.4  0.8075  0.8251  0.9456  1.0000
```

Based on the above correlations, it is possible an AR(1) model is appropriate because the correlations between the time points 1 apart were 0.941, 0.918, and 0.946, which are similar.

```
#fit AR1 model
ar1 <- gee(ht ~ age_yr, id = boy, corstr = 'AR-M', data = rao)
```

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

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

```
## (Intercept)      age_yr
##     34.3625      1.8500
```

```
summary(ar1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     AR-M , M = 1
##
## Call:
## gee(formula = ht ~ age_yr, id = boy, data = rao, corstr = "AR-M")
##
## Summary of Residuals:
##       Min        1Q    Median        3Q       Max
## -5.208555 -2.208555 -0.208555  2.220042  5.791445
##
##
## Coefficients:
##               Estimate Naive S.E.  Naive z Robust S.E.  Robust z
## (Intercept) 44.1362601  3.4564524 12.76924   1.9149505 23.048251
## age_yr       0.7143876  0.4101479  1.74178   0.2194097  3.255953
##
## Estimated Scale Parameter:  7.746551
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.8431426 0.7108894 0.5993811
## [2,] 0.8431426 1.0000000 0.8431426 0.7108894
## [3,] 0.7108894 0.8431426 1.0000000 0.8431426
## [4,] 0.5993811 0.7108894 0.8431426 1.0000000
```

```
#get confidence intervals
#age coefficient +/- 1.96*std error
lb_ar <- ar1$coefficients[2] - 1.96*0.219
ub_ar <- ar1$coefficients[2] + 1.96*0.219
paste("Estimated bone growth per year is", round(ar1$coefficients[2],3), 'mm.')
```

```
## [1] "Estimated bone growth per year is 0.714 mm."
```

```
paste("AR1 95% CI = (",round(lb_ar,3),',',round(ub_ar,3),')')
```

```
## [1] "AR1 95% CI = ( 0.285 , 1.144 )"
```

Part b) [BONUS] Estimate the rate of bone growth per year using a random effects model (using a single random effect only). Give a 95% CI for the rate. Do you think this inference is valid? Explain in detail.

```
library(lme4)
```

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

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

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

```
rand_eff <- lmer(ht ~ age_yr + (1|boy), data = rao)
summary(rand_eff)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ht ~ age_yr + (1 | boy)
##    Data: rao
##
## REML criterion at convergence: 321.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0226 -0.5208 -0.0939  0.4368  3.3402
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  boy      (Intercept) 6.017    2.453
##  Residual             1.704    1.305
## Number of obs: 80, groups:  boy, 20
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  34.3625     3.5542   9.668
## age_yr        1.8500     0.4128   4.482
##
## Correlation of Fixed Effects:
##        (Intr)
## age_yr -0.987
```

```
ints <- confint(rand_eff)[4,]
```

```
## Computing profile confidence intervals ...
```

```
paste("Estimated bone growth per year is", 1.850, 'mm.')
```

```
## [1] "Estimated bone growth per year is 1.85 mm."
```

```
paste("95% CI = (",round(ints[1],3),',',round(ints[2],3),')')
```

```
## [1] "95% CI = ( 1.035 , 2.665 )"
```

This model assumes independent variance, so the inference is likely invalid.

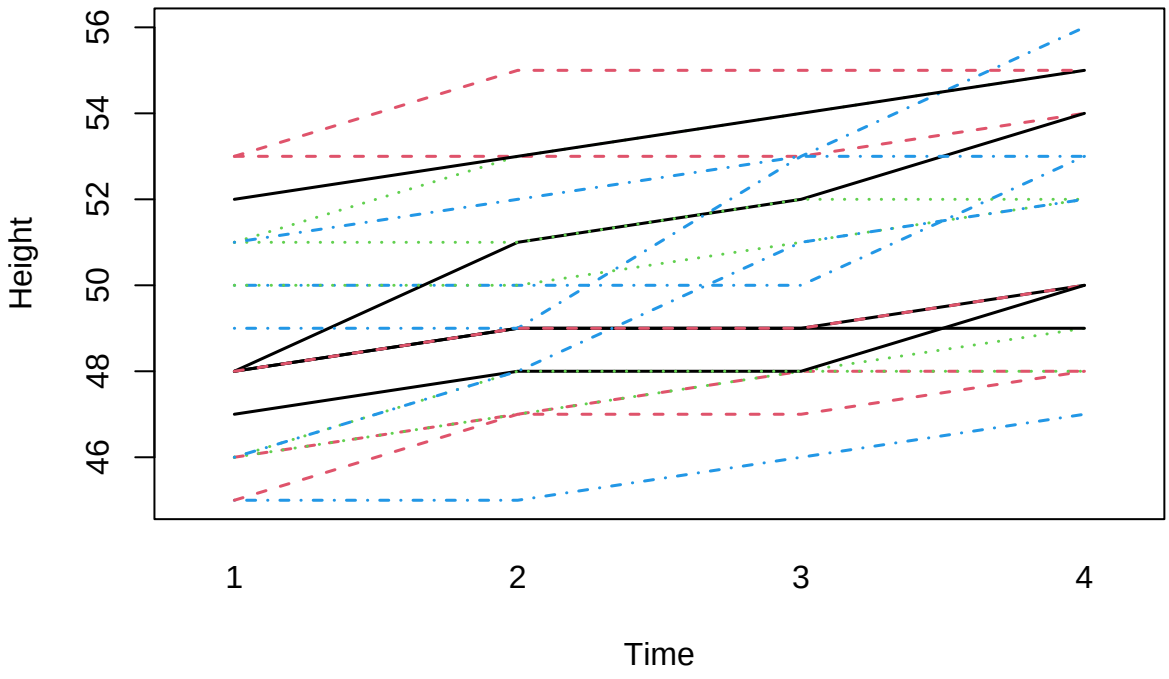
Part c) Estimate the rate of bone growth per year for each boy separately (hint: use a derived variable). Calculate the mean and standard deviation of the rates. Suggest a method for using these results to compute a 95% CI for the overall rate and compute this interval.

```
#plot ramus height over time
attach(rao)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     mnht, sdht
```

```
## The following objects are masked from rao (pos = 5):
##
##     age, age_yr, boy, ht, mnht, sdht, stdht
```

```
interaction.plot(age,boy,ht,legend = F, col = 1:4,lty=1:4, xlab = 'Time',ylab= 'Height', lwd = 1.5)
```



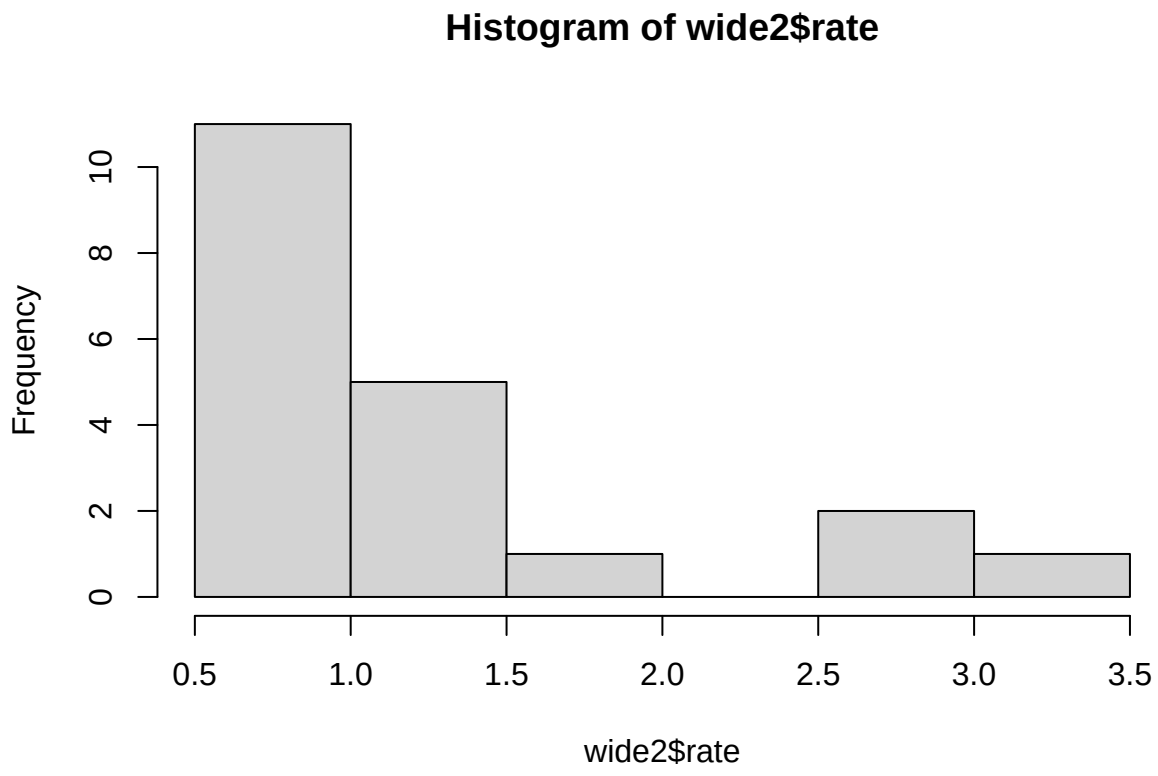
```r
#derive variable for avg change in ramus height for each boy
rao.sort2 <- rao[order(rao$boy),][,c(1,2,3)]
wide2 <- reshape(rao.sort2,timevar = "age",idvar = c("boy"),direction="wide")

wide2$rate <- (wide2$ht.4 - wide2$ht.1)/2

#get mean and sd
paste('Mean of Average Change in Ramus Height per Year =',mean(wide2$rate))
```

```
## [1] "Mean of Average Change in Ramus Height per Year = 1.425"
```

```r
paste('SD of Average Change in Ramus Height per Year =',round(sd(wide2$rate),3))
```

```
## [1] "SD of Average Change in Ramus Height per Year = 0.847"
```

```r
hist(wide2$rate)
```




```r
t.test(wide2$rate)
```

```
##
##  One Sample t-test
##
## data:  wide2$rate
```

```
## t = 7.5221, df = 19, p-value = 4.132e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.028494 1.821506
## sample estimates:
## mean of x
##     1.425
```

Based on the results of the t-test, the estimated average rate of growth (mm/year) is 1.42 with a 95% CI = (1.03,1.82).

Part d) Compare the results in (a), (b), and (c), and explain the similarities and differences between them.

Part a Independence 95% CI = (1.291 , 2.409), Exchangeable 95% CI = (1.291 , 2.409), AR1 95% CI = (0.285 , 1.144)

Part b 95% CI = (1.035, 2.665)

Part c 95% CI = (1.03,1.82)

The AR1 model had the most different results. This is likely because it specifies a variance structure that is very incorrect for this data set. The independent and exchangeable models in part a gave the exact same answers because of the robust standard errors from the sandwich estimators. Part b gave a wider confidence interval because it did not use a robust estimate. Part c gave an answer similar to part b but with a narrower confidence interval because it is making more strict assumptions for the t-test, which is likely too conservative of an estimate. Overall, I would probably trust the independent model in part a the most because it relaxes the assumptions made in part b and accounts for the correlation while using all of the data (unlike part c).

Question 2: This exercise involves an analysis of the data from a chemotherapy trial for seizures in epileptic patients. The data description is below. The data set, epilepsy.xls, is available on the Sakai class website.

Data are from Thall and Vail, Biometrics 46:657-671, 1990

Results of a randomized clinical trial of Progabide for treatment of patients with epilepsy (in wide format). Data on N=59 patient had reported seizures during four consecutive two-week periods following randomization. Baseline seizure counts during the eight-week period prior to treatment assignment were also collected.

Var Name Description Range/Values 1 id Patient ID 104-236 2 y1 Seizure count in 1st period 0-102 3 y2 Seizure count in 2nd period 0-65 4 y3 Seizure count in 3rd period 0-76 5 y4 Seizure count in 4th period 0-63 6 tx Treatment assignment 0 = Control, 1 = Progabide 7 y0 Seizure count, 8-week 0-151 baseline period 8 age Age (in years) at baseline 18-42

```
library(readxl)
epilepsy <- read_excel("C:/Users/sdm98/Desktop/BIOSTAT718/epilepsy.xls")
```

Part a) Estimate and test the significance of the treatment effect on the average seizure rate after randomization. Use a Poisson regression model with adjustment for the baseline seizure rate and age. (Please use a transformation of the baseline count that seems appropriate for this model.) Do this analysis two ways: (1) using naïve variance estimates, and (2) using robust variance estimates. Explain the difference between the naïve variance and robust variance estimates in this analysis.

run gee using independent correlation to get naive and robust

```r
#transform y0 to be rate for a two-week period to match other rates
epilepsy$y0_t <- epilepsy$y0/4

#convert from wide to long
epilepsy_l <- gather(epilepsy, y, count, y1:y4, factor_key = TRUE)
epilepsy_l
```

```
## # A tibble: 236 x 7
##       id    rx    y0   age  y0_t y     count
##    <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <dbl>
##  1   104     0    11    31  2.75 y1        5
##  2   106     0    11    30  2.75 y1        3
##  3   107     0     6    25  1.5  y1        2
##  4   114     0     8    36  2    y1        4
##  5   116     0    66    22 16.5  y1        7
##  6   118     0    27    29  6.75 y1        5
##  7   123     0    12    31  3    y1        6
##  8   126     0    52    42 13    y1       40
##  9   130     0    23    37  5.75 y1        5
## 10   135     0    10    28  2.5  y1       14
## # ... with 226 more rows
```

```r
indep2 <- gee(count ~ rx + age + y0_t, family = 'poisson', id = id, corstr = 'independence', data = epi
```

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

## running glm to get initial regression estimate

## (Intercept)          rx         age        y0_t
##  0.56253157 -0.15270095  0.02274013  0.09060695
```

```r
summary(indep2)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Independent
##
## Call:
## gee(formula = count ~ rx + age + y0_t, id = id, data = epilepsy_l,
##     family = "poisson", corstr = "independence")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -15.895799  -3.341317  -1.195079   1.318402  63.931712
##
##
```

```
## Coefficients:
##                Estimate  Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept)  0.56253157 0.305958882  1.838586 0.262228811  2.145194
## rx          -0.15270095 0.107849076 -1.415876 0.113904570 -1.340604
## age          0.02274013 0.009078201  2.504916 0.008182567  2.779096
## y0_t         0.09060695 0.004595965 19.714456 0.003819201 23.724061
##
## Estimated Scale Parameter:  5.089608
## Number of Iterations:  1
##
## Working Correlation
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    0    0    0
## [3,]    0    0    0    0
## [4,]    0    0    0    0
```

```
#get confidence intervals
#naive
lb_indep2 <- indep2$coefficients[2] - 1.96*0.10785
ub_indep2 <- indep2$coefficients[2] + 1.96*0.10785
paste("The growth rate for those on the treatment is estimated to be", round(exp(indep2$coefficients[2]]
```

```
## [1] "The growth rate for those on the treatment is estimated to be 0.858 times the seizure rate for t
```

```
paste("Naive 95% CI = (",round(exp(lb_indep2),3),',',round(exp(ub_indep2),3),')')
```

```
## [1] "Naive 95% CI = ( 0.695 , 1.06 )"
```

```
#robust
lb_indep2_r <- indep2$coefficients[2] - 1.96*0.11390
ub_indep2_r <- indep2$coefficients[2] + 1.96*0.11390
paste("Robust 95% CI = (",round(exp(lb_indep2_r),3),',',round(exp(ub_indep2_r),3),')')
```

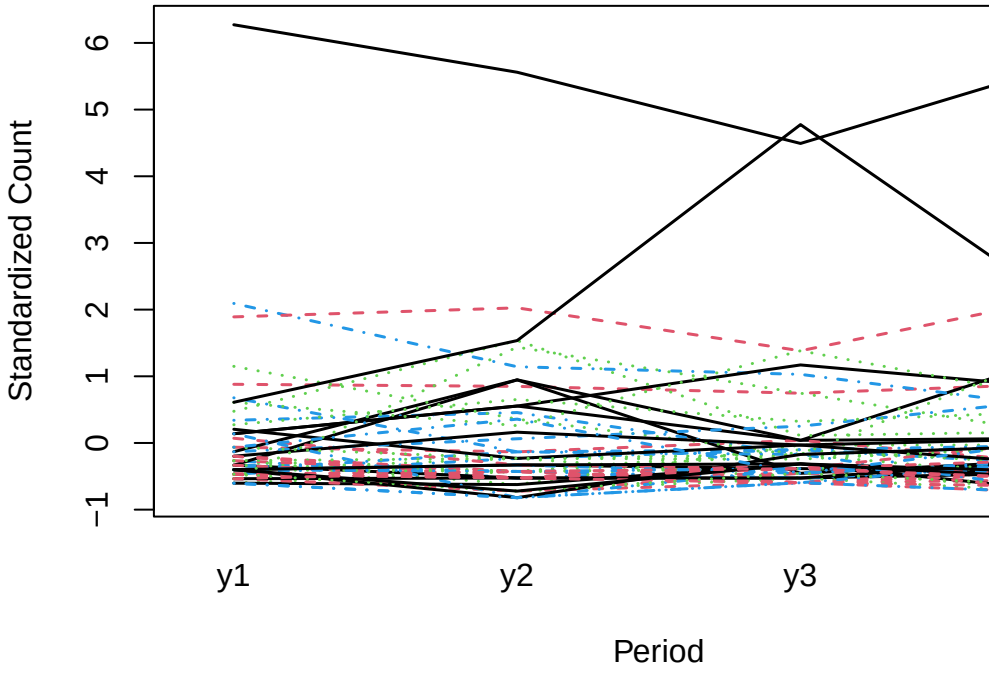
```
## [1] "Robust 95% CI = ( 0.687 , 1.073 )"
```

The naive variance estimate is generated using the poisson model. The robust variance estimate is generated such that the poisson assumption is more relaxed. This results in a wider confidence interval.

Part b) Estimate and test the significance of the treatment effect using a GEE analysis of the individual, 2-week seizure counts. Use a Poisson-like regression model with a working correlation matrix that seems appropriate for these data (justify your choice), and adjust for baseline seizure rate and age as in part (a).

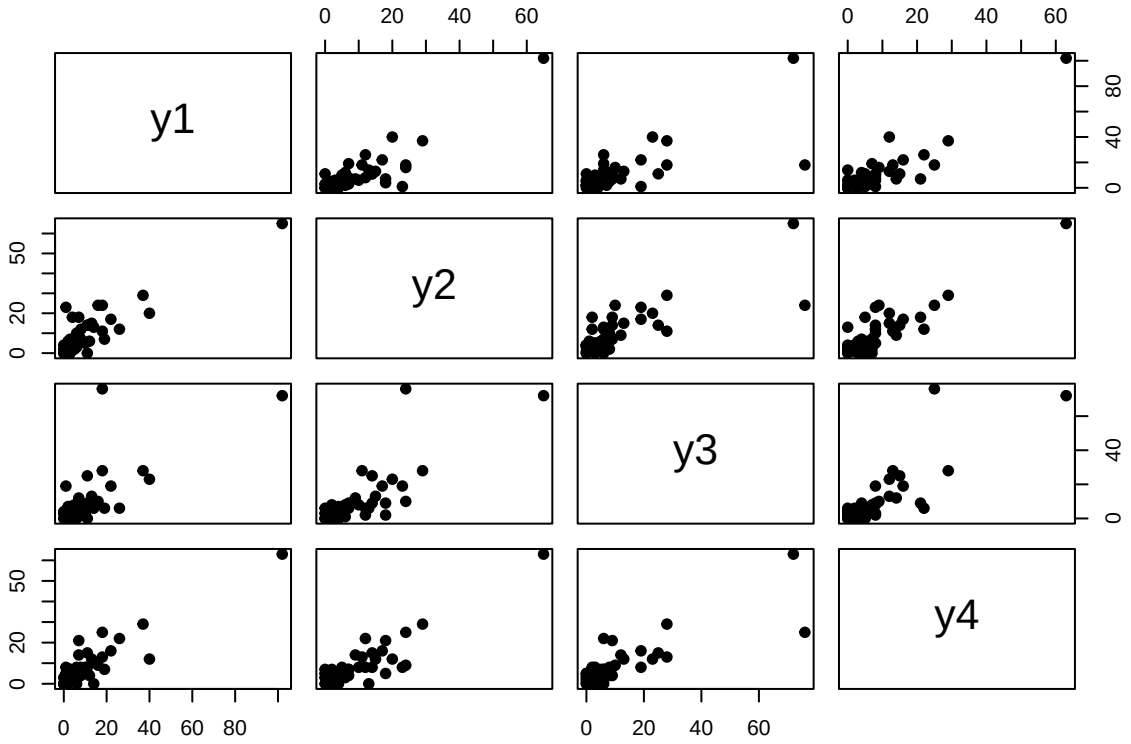
```
#first, look at the change in count over time for each patient
mnct <- with(epilepsy_l, tapply(count,y,mean))
epilepsy_l$mnct <- mnct[match(epilepsy_l$y,names(mnct))]
sdct <- with(epilepsy_l, tapply(count, y, sd))
```

```
epilepsy_1$sdct <- sdct[match(epilepsy_1$y,names(sdct))]
epilepsy_1$sdct <- (epilepsy_1$count - epilepsy_1$mnct)/epilepsy_1$sdct

interaction.plot(epilepsy_1$y,epilepsy_1$id,epilepsy_1$sdct,legend = F, col = 1:4,lty=1:4, xlab = 'Peri
```



Determine correlation structure

```
#next, examine residuals

#fit independent model to get residuals
mod2 <- lm(count~y, data = epilepsy_1)
epilepsy_1$res <- mod2$residuals

#plot scatter plots of residuals by time points
pairs(epilepsy[,2:5],pch=19)
```




```
#get correlation table
round(cor(epilepsy[,2:5]),4)
```

```
##        y1     y2     y3     y4
## y1 1.0000 0.8708 0.7377 0.8925
## y2 0.8708 1.0000 0.8025 0.8951
## y3 0.7377 0.8025 1.0000 0.8240
## y4 0.8925 0.8951 0.8240 1.0000
```

This suggests an unstructured correlation is most appropriate.

```
#run unstructured gee model
unstr2 <- gee(count ~ rx + age + y0_t, family = 'poisson', id = id, corstr = 'unstructured', data = epi
```

Run Model

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

## running glm to get initial regression estimate

## (Intercept)          rx         age        y0_t
##  0.56253157 -0.15270095  0.02274013  0.09060695
```

```
summary(unstr2)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = count ~ rx + age + y0_t, id = id, data = epilepsy_l,
##     family = "poisson", corstr = "unstructured")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -15.895799  -3.341317  -1.195079   1.318402  63.931712
##
##
## Coefficients:
##                Estimate  Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept)  0.56253157 0.305958882  1.838586 0.262228811  2.145194
## rx          -0.15270095 0.107849076 -1.415876 0.113904570 -1.340604
## age          0.02274013 0.009078201  2.504916 0.008182567  2.779096
## y0_t         0.09060695 0.004595965 19.714456 0.003819201 23.724061
##
## Estimated Scale Parameter:  5.089608
## Number of Iterations:  1
##
## Working Correlation
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    0    0    0
## [3,]    0    0    0    0
## [4,]    0    0    0    0
```

```
#get confidence intervals
lb3 <- unstr2$coefficients[2] - 1.96*0.11390
ub3 <- unstr2$coefficients[2] + 1.96*0.11390
paste("The seizure rate for those on the treatment is estimated to be", round(exp(unstr2$coefficients[2]
```

```
## [1] "The seizure rate for those on the treatment is estimated to be 0.858 times the seizure rate for
```

```
paste("95% CI = (",round(exp(lb3),3),',',round(exp(ub3),3),')')
```

```
## [1] "95% CI = ( 0.687 , 1.073 )"
```

This is a non-significant result, meaning there is no evidence of a treatment effect on seizure rate after controlling for baseline age and baseline seizure rate.

Part c) Compare the results on the treatment effect for (a) and (b). Are there substantial differences in the inference about the treatment effect? If so, give possible reasons for these differences. What method do you feel is best for performing inference about the treatment effect (please give reasons)?

Estimated average effect = 0.858 for all models.

Part a Naive 95% CI = (0.695, 1.06)

Part a Robust 95% CI = (0.687, 1.073)

Part b Robust 95% CI = (0.687, 1.073)

The robust estimates gave the same results, so working independence with a robust estimator would be the best (simplest) choice. Robust is prefered to the naive estimate because it is relaxing the assumptions of the model to better estimate the true patterns in the data.