BIS 628 HW 6

Joanna Chen

4/20/2020

Question 1 (25 points):

(a) (4 points) Go back to HW 4 – is the study design, which produced the data, suitable for fitting a marginal model via GEE? Why or why not?

It's not suitable. Because small sample size and multiple time points. Therefore it's not suitable for GEE. GEE wants to use sandwich estimator for statistical inference and usually requires large sample size. If MAR or small sample size, we can consider use model-based approachs.

(b) (5 points) On slide 13 of Lecture 10, what are the implications of this statement: "If estimated $V_i = (Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)'$, then the two estimators (naive or model-based vs. empirical or sandwich) are the same." What does it mean that the naïve or model-based estimator is the same as the sandwich estimator?

That means 1. our chosen model is great. 2. We discovered the true correlation here. Also books says that

$$\text{If we model } V_i \text{ correctly, } V_i \equiv \Sigma_i \text{ , and } \text{Cov}(\widehat{\beta}) = \text{B}^{-1}.$$

(c) (6 points) While there is no guideline that would tell you which analytical approach, a GLMM vs. a marginal model via GEE, is better. What would be 3 advantages of implementing a GEE model over a GLMM model?

We can find the answer on the book Chapter 13 Page 6. First, the GEE estimator $\hat{\beta}$ is almost as precise or efficient as the MLE. As a result, there is little loss of precision when the GEE approch is adopted as an alternative to maximum likehood. Second, the GEE estimator $\hat{\beta}$ is a consistent estimator of β even if the within-subject associations among the repeated measures have been misspecified; this is a very appealing robustness property of the GEE estimator. Third, GEE approach can be applied equally to continuous responses.

(d) (10 points) If you were to extend a generalized linear model for a binary response to longitudinal data, and assuming that there exists a very computentially intensive approach to get marginal probabilities from the GLMM, how would you go about making equivalent GEE and GLMM models? Give a specific example of such two approximately 'equivalent' models. Hint: You should consider all three specification for a GLM model, the within-subject dependency, the conditional assumption of independence for Yij, and the assumptions for the types of covariates. In order to get the full points, you should consider all of these.

For GLMM, we want to make subject-level inference. We describe the covariate effect on a typical subject's response. The process involves estimate β^* to predict b_i using ML. The assumption of missing values are MAR. For GLMM, we want to make population-level inference. We assume $\text{Cov}(Y) = V$ to estimate β then use Pearson residuals to estimate the component of V and repeat this process. We need to choose a working covariance here and select the best covariance which is has the least QIC. The assumption of missing values are MCAR. If we only have a random intercept, setting the working correlation as CS or exchangable, then it is appoximatly equivalent. In that way we may show that the marginalized estimate is about the same. Regarding the assumptions for the types of covariates, GLMM more supports time-varing covariates. GEE is for time-invariant covariates. This is on slides page 10 and book 13.5.

Question 2 (55 points):

The data set of smoking cessation trial is posted on the course webpage. Participants were randomized to either control, discussion or social support condition. Some subjects that were randomized to one of the two treatments (discussion or social support) never showed up for the discussion or social support gathering and these subjects were labeled as "noshow" in the data set. The data were collected at four telephone interviews (occasion): 0, 6, 12 and 24 months post- intervention (month). The longitudinal binary response variable is smoking abstinence (quit). Additional covariate of importance is race (white vs. nonwhite). Dataset: datasmoking.csv or datasmoking.xlsx
The dataset is already in a long format, i.e. with repeated observations within a person. The variables are (in column order):
Variable List:
ID: participant ID number
Quit (binary outcome): 0=still smoking, 1=stopped smoking
Occasion: visit number, starting from 0: 0=baseline, 1=6 months, 2=12 months, 3=24 months. Racew: a dummy variable, indicating 1=White, 0=Not White
Dummy coded Group variables: control, noshow, discussion, socsup
Month: 0,6,12,24 months

Implement a marginal model for the correlated outcome, Quit (smoking cessation), using a GEE analysis, including the following variables: Group, Racew, time (Month and Month-squared) and the interaction terms between White race and time, as well as Group and time.

```r
setwd("~/Downloads/HW 6")
library(readr)
datasmoking <- read_csv("datasmoking.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   quit = col_double(),
##   occasion = col_double(),
##   racew = col_double(),
##   control = col_double(),
##   noshow = col_double(),
##   discussion = col_double(),
##   socsup = col_double(),
##   month = col_double()
## )
```

```r
### Fit marginal models by the GEE method
#install.packages("geepack")
library(geepack)
datasmoking$month2 = (datasmoking$month)^2
m1.ind <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew*mon
summary(m1.ind)
```

```
##
## Call:
## geeglm(formula = quit ~ control + discussion + socsup + racew +
##     month + month2 + racew * month + racew * month2 + control *
##     month + control * month2 + discussion * month + discussion *
##     month2 + socsup * month + socsup * month2, family = binomial(link = "logit"),
##     data = datasmoking, id = id, corstr = "independence", std.err = "san.se")
##
##  Coefficients:
```

```
##                      Estimate    Std.err   Wald Pr(>|W|)    
## (Intercept)       -1.0336377  0.1830837 31.874 1.64e-08 ***
## control           -0.6789085  0.3128452  4.709 0.029999 *  
## discussion         0.3021548  0.2854675  1.120 0.289848    
## socsup             0.8998297  0.2567039 12.287 0.000456 ***
## racew              0.0186560  0.2508195  0.006 0.940708    
## month             -0.1108454  0.0381019  8.463 0.003624 ** 
## month2             0.0035785  0.0014250  6.306 0.012034 *  
## racew:month        0.1091607  0.0464536  5.522 0.018779 *  
## racew:month2      -0.0036835  0.0017781  4.292 0.038298 *  
## control:month      0.0365467  0.0562378  0.422 0.515783    
## control:month2    -0.0003315  0.0021179  0.025 0.875611    
## discussion:month  -0.0650166  0.0628472  1.070 0.300894    
## discussion:month2  0.0028980  0.0024084  1.448 0.228855    
## socsup:month      -0.0925307  0.0550911  2.821 0.093036 .  
## socsup:month2      0.0032316  0.0020975  2.374 0.123393    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Correlation structure = independence 
## Estimated Scale Parameters:
## 
##             Estimate Std.err
## (Intercept)    1.003 0.09725
## Number of clusters:   489  Maximum cluster size: 4
```

(a) 4 points: Write out the equation for this marginal model.

$$
\begin{aligned}
&\log\left\{\Pr\left(Y_{ij}=1|X_{ij}\right)/\Pr\left(Y_{ij}=0|X_{ij}\right)\right\}\\
=&\beta_1+\beta_2\text{Ctrl}_i+\beta_3\text{Discussion}_i+\beta_4\text{Socsup}_i+\beta_5\text{Racew}_i+\beta_6\text{Month}_{ij}+\beta_7\text{Month}^2_{ij}+\beta_8\text{Racew}_i\text{Month}_{ij}+\beta_9\text{Racew}_i\text{Month}^2_{ij}\\
&+\beta_{11}\text{Ctrl}_i\text{Month}^2_{ij}+\beta_{12}\text{Discussion}_i\text{Month}_{ij}+\beta_{13}\text{Discussion}_i\text{Month}^2_{ij}+\beta_{14}\text{Socsup}_i\text{Month}_{ij}+\beta_{15}\text{Socsup}_i\text{Month}^2_{ij}
\end{aligned}
$$

(b) 2 points: What important 'conditional' assumption for the marginal mean is in this model?

Mean response is conditional only on covariates at that time and NOT on other responses or covariates at other time points and NOT on random effects. Previous time does not effect current time.

(c) 2 points:What is the link function?

$$\text{Canonical Link function for mean response, } \mu_{ij} : g\left(\mu_{ij}\right)=\log\left(\mu_i/\left(1-\mu_{ij}\right)\right.$$

(d) Using the sandwich estimator, and the independence assumption for the within-subject association (working correlation), report what you find in terms of covariates effects on smoking cessation and test using generalized Wald statistics for the statistical significance at alpha level of 0.05 for each interpretable grouping of interaction terms (this means that some interaction terms have to be tested together).

(i) 5 points: Write out the null and alternative hypotheses for the interaction of White race and time. State your conclusion based on the statistical test and write-out in words your interpretation.

H0: Coefficient of Race * Month = Coefficient of Race * Month2 = 0

$$H_0 : \beta_8 = \beta_9 = 0, H_1 : \text{At least one of betas is nonzero.}$$

```
m1.ind.test1 <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + control*month + c
anova(m1.ind,m1.ind.test1)
```

```
## Analysis of 'Wald statistic' Table
##
## Model 1 quit ~ control + discussion + socsup + racew + month + month2 + racew * month + racew * mont
## Model 2 quit ~ control + discussion + socsup + racew + month + month2 + control * month + control *
##   Df   X2 P(>|Chi|)
## 1  2 5.89     0.053 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The conclusion p = 0.053 > 0.05 therefore we fail to reject the null hypothesis and conclude that the interaction of White race and time are not significant.

(ii) 5 points: Write out the null and alternative hypotheses for the interaction of Group and time. State your conclusion based on the statistical test and write-out in words your interpretation.

H_0: Coefficient of Socsup * Month = Coefficient of Ctrl * Month = Coefficient of Discussion * Month = Coefficient of Socsup * Month2 = Coefficient of Ctrl * Month2 = Coefficient of Discussion * Month2 = 0

$$H_0 : \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0, H_1 : \text{At least one of betas is nonzero.}$$

```
m1.ind.test2 <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + rac
anova(m1.ind,m1.ind.test2)
```

```
## Analysis of 'Wald statistic' Table
##
## Model 1 quit ~ control + discussion + socsup + racew + month + month2 + racew * month + racew * mont
## Model 2 quit ~ control + discussion + socsup + racew + month + month2 + racew * month + racew * mont
##   Df   X2 P(>|Chi|)
## 1  6 8.51       0.2
```

The conclusion p = 0.2 > 0.05 therefore we fail to reject the null hypothesis and conclude that the interaction of time and group are not significant.

(e) 4 points: Plot the estimated marginal probabilities of smoking cessation over time from the final model in Q2(d):

Final model:

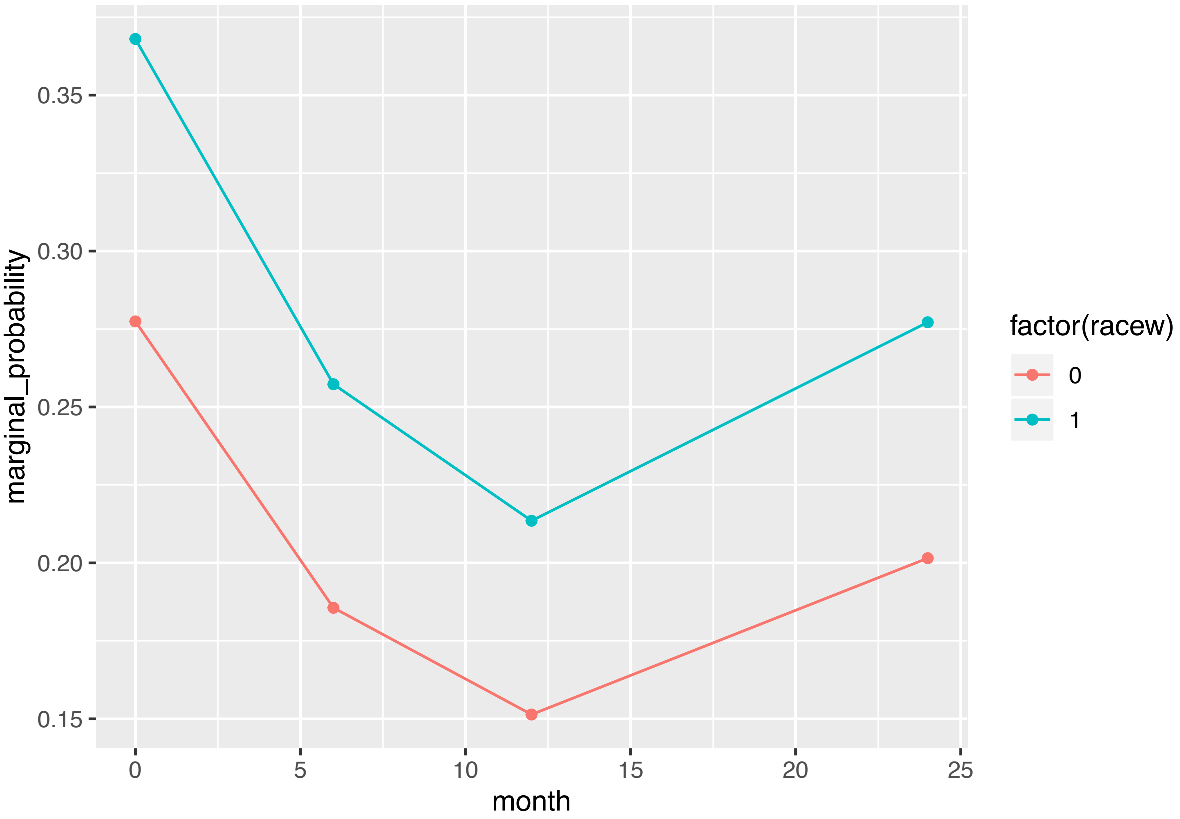
```
m2 <- geeglm(quit ~ control + discussion + socsup +racew + month + month2, data=datasmoking, family=bin

m2_pred <- unique(data.frame(control=datasmoking$control, discussion=datasmoking$discussion, socsup=data
#exp(fitted(m2))/(1+exp(fitted(m2)?
#Print out m2_pred
print(m2_pred)
```

```
##     control discussion socsup racew month month2 fitted.m2.
## 1         0          1      0     0     0      0    0.29155
## 2         0          1      0     0     6     36    0.19455
## 3         0          1      0     0    12    144    0.15846
## 4         0          1      0     0    24    576    0.21135
## 5         0          0      0     0     0      0    0.25362
## 6         0          0      0     0     6     36    0.16628
## 7         0          0      0     0    12    144    0.13455
## 8         0          0      0     0    24    576    0.18118
## 36        0          0      0     1     0      0    0.34255
## 37        0          0      0     1     6     36    0.23420
## 38        0          0      0     1    12    144    0.19250
## 39        0          0      0     1    24    576    0.25334
## 40        0          0      1     1     0      0    0.48240
## 44        0          0      1     0     0      0    0.37804
## 45        0          0      1     0     6     36    0.26295
## 46        0          0      1     0    12    144    0.21759
## 47        0          0      1     0    24    576    0.28357
## 106       0          0      1     1     6     36    0.35361
## 107       0          0      1     1    12    144    0.29895
## 152       1          0      0     1     0      0    0.26018
## 153       1          0      0     1     6     36    0.17110
## 154       1          0      0     1    12    144    0.13861
## 155       1          0      0     1    24    576    0.18634
## 213       1          0      0     0     0      0    0.18656
## 214       1          0      0     0     6     36    0.11865
## 215       1          0      0     0    12    144    0.09497
## 216       1          0      0     0    24    576    0.12995
## 313       0          1      0     1     0      0    0.38689
## 314       0          1      0     1     6     36    0.27028
## 315       0          1      0     1    12    144    0.22404
## 316       0          1      0     1    24    576    0.29124
## 388       0          0      1     1    24    576    0.37769
```

a. By White vs. Non-White Race, and comment on what you see.

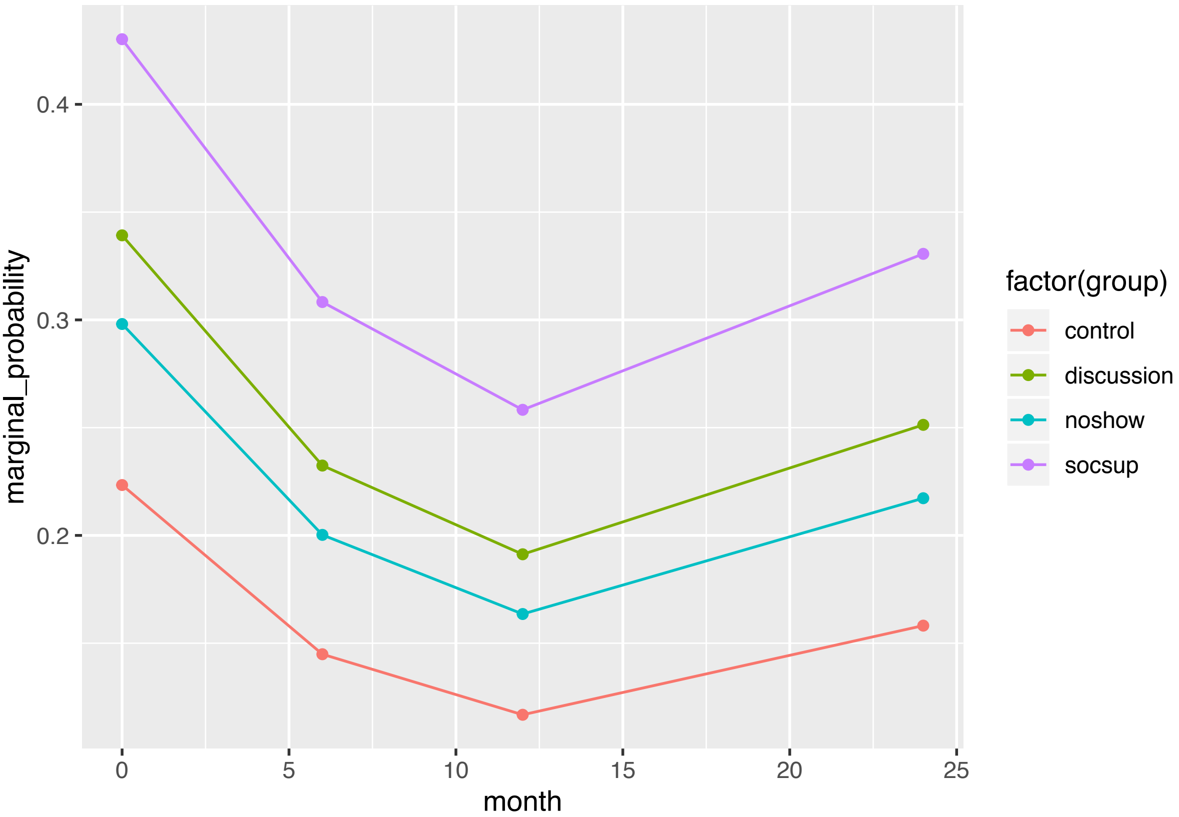
```
library(data.table)
library(ggplot2)
plot_t1 = setDT(m2_pred)
setnames(plot_t1,"fitted.m2.","marginal_probability")
plot_t1 = plot_t1[,c("racew","month","marginal_probability")]
setDT(plot_t1)
plot_t1 = plot_t1[,mean(marginal_probability),by=c("racew","month")]
setnames(plot_t1,"V1","marginal_probability")
ggplot(plot_t1,aes(x=month,y = marginal_probability, color = factor(racew),group = factor(racew))) + ge
```



White race have higher probability to stop smoking than the non-white race.

b. By Group, and comment on what you see.

```
plot_t2 = setDT(m2_pred)
plot_t2 = plot_t2[,c("control","discussion","socsup","month","marginal_probability")]
setDT(plot_t2)
plot_t2 = plot_t2[,mean(marginal_probability),by=c("control","discussion","socsup","month")]
setnames(plot_t2,"V1","marginal_probability")
plot_t2 = plot_t2[, group:=ifelse(discussion == 1,"discussion",ifelse(socsup ==1,"socsup",ifelse(control
ggplot(plot_t2,aes(x=month,y = marginal_probability, color = factor(group),group = factor(group))) + geo
```



The group with social support has higher marginal probability to stop smoking then the discussion group, follows with the group with the third high probability noshow group. The control group has the lowest marginal probability to stop smoking.

(f) 5 Points: Starting with the full marginal model again (fixed effects :Group,Racew,time (Month and Month-squared) and the interaction terms between White race and time, as well as Group and time), use QIC to select a "best" working correlation structure among independent, unstructured (fullcluster on the logodds level), exchangeable (on the log-odds level), Toeplitz (on the log-odds level), and AR(1). Once you select the 'best' working correlation, interpret your estimates for the within-subject association.

```
#1. working correlation structure: Independent
m3.ind <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew*mon

#2. working correlation structure: Unstructured
m3.un <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew*mont
                 id=id, corstr = "unstructured", std.err="san.se")
#summary(m1.un)

#3. working correlation structure: Exchangeable
m3.exch <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew*mo
                 id=id, corstr = "exchangeable", std.err="san.se")
#summary(m1.exch)

#4. working correlation structure: Toeplitz
#sigma_ij = sigma_|i-j|
set.seed(123)
```

```
#generating the design matrix for the unstructured correlation
zcor <- genZcor(clusz = table(datasmoking$id), waves = datasmoking$occasion+1, corstrv=4)
# defining the Toeplitz structure
zcor.toep<-matrix(NA, nrow(zcor),3)
zcor.toep[,1]<-apply(zcor[,c(1,4,6)],1,sum)
zcor.toep[,2]<-apply(zcor[,c(2,5)],1,sum)
zcor.toep[,3]<-zcor[,3]

m3.toep <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew*mon

#5. working correlation structure: AR(1)
m3.ar1 <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew*mont
                 id=id, corstr = "ar1", std.err="san.se")
#summary(m3.ar1)

m3.exch <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew*mon
                 id=id, corstr = "toeplitz", std.err="san.se")
```

```
## Warning in if (corstrv == -1) stop("invalid corstr."): the condition has length
## > 1 and only the first element will be used
```

```
## Warning in if (corstrv == 5) stop("need zcor matrix for userdefined corstr.")
## else zcor <- genZcor(clusz, : the condition has length > 1 and only the first
## element will be used
```

```
## Warning in if (corstrv == 1) return(matrix(0, 0, 0)): the condition has length >
## 1 and only the first element will be used
```

```
## Warning in if (corstrv == 6) alpha <- 1 else alpha <- rep(0, q): the condition
## has length > 1 and only the first element will be used
```

```
#OBTAIN QIC to compare models: takes geeglm function as the input
QIC(m3.ar1)
```

```
##        QIC      QICu Quasi Lik       CIC    params      QICC
##     1792.7    1792.9    -881.4      14.9      15.0    1793.0
```

```
QIC(m3.exch)
```

```
##        QIC      QICu Quasi Lik       CIC    params      QICC
##    1793.90   1792.40   -881.20     15.75     15.00   1794.21
```

```
QIC(m3.un)
```

```
##        QIC      QICu Quasi Lik       CIC    params      QICC
##    1793.42   1792.58   -881.29     15.42     15.00   1793.96
```

```
QIC(m3.ind)
```

```
##        QIC      QICu Quasi Lik       CIC    params      QICC
##    1796.90   1792.13   -881.07     17.38     15.00   1797.17
```

```
QIC(m3.toep)
```

```
##        QIC      QICu Quasi Lik       CIC    params      QICC
##    1792.37   1792.34   -881.17     15.01     15.00   1792.76
```

```
summary(m3.toep)
```

```
##
## Call:
## geeglm(formula = quit ~ control + discussion + socsup + racew +
##     month + month2 + racew * month + racew * month2 + control *
##     month + control * month2 + discussion * month + discussion *
##     month2 + socsup * month + socsup * month2, family = binomial(link = "logit"),
##     data = datasmoking, id = id, zcor = zcor.toep, corstr = "userdefined",
##     std.err = "san.se")
##
##  Coefficients:
##                     Estimate   Std.err  Wald Pr(>|W|)
## (Intercept)        -1.002759  0.182328 30.25  3.8e-08 ***
## control            -0.674218  0.309729  4.74  0.02950 *
## discussion          0.287353  0.286552  1.01  0.31596
## socsup              0.886946  0.255925 12.01  0.00053 ***
## racew              -0.003872  0.249585  0.00  0.98762
## month              -0.111511  0.037187  8.99  0.00271 **
## month2              0.003586  0.001386  6.69  0.00967 **
## racew:month         0.102338  0.045825  4.99  0.02553 *
## racew:month2       -0.003250  0.001734  3.51  0.06087 .
## control:month       0.032390  0.054838  0.35  0.55475
## control:month2     -0.000214  0.002048  0.01  0.91693
## discussion:month   -0.062726  0.061316  1.05  0.30631
## discussion:month2   0.002783  0.002332  1.42  0.23269
## socsup:month       -0.093701  0.054335  2.97  0.08462 .
## socsup:month2       0.003282  0.002062  2.53  0.11155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = userdefined
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.996  0.0944
##   Link = identity
##
## Estimated Correlation Parameters:
##         Estimate Std.err
## alpha:1    0.409  0.0574
## alpha:2    0.306  0.0553
## alpha:3    0.251  0.0618
## Number of clusters:   489  Maximum cluster size: 4
```

By comparing QIC, Toeplitz gives the best working correlation structure.
Interpret your estimates for the within-subject association:
alpha1: the pairwise log odds ratios between occasions whose difference is 1 is 0.409.
alpha2: the pairwise log odds ratios between occasions whose difference is 2 is 0.306.
alpha3: the pairwise log odds ratios between occasions whose difference is 3 is 0.251.

(g) Refit your full marginal model, but now using the 'best' working correlation structure found in (e), and test the covariates of interest.

```
m3.toep <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew*mon
```

(i) 5 points: Following the instructions for Q2(d)(i), what is your inference for the effect of White race on the probability of smoking cessation over time using the sandwich estimator with the ‘best’ working correlation from Q2(f)?

H0: Coefficient of Race * Month = Coefficient of Race * Month^2 = 0

$$H_0 : \beta_8 = \beta_9 = 0, H_1 : \text{At least one of betas is nonzero.}$$

```r
m3.toep.test1 <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + control*month +

anova(m3.toep,m3.toep.test1)
```

```
## Analysis of 'Wald statistic' Table
##
## Model 1 quit ~ control + discussion + socsup + racew + month + month2 + racew * month + racew * mont
## Model 2 quit ~ control + discussion + socsup + racew + month + month2 + control * month + control * 
##   Df   X2 P(>|Chi|)
## 1  2 5.96     0.051 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The conclusion p = 0.051 > 0.05 therefore we fail to reject the null hypothesis and conclude that the interaction of White race and time are not significant.

(ii) 5 points: Similarly, what is your inference for the effect of White race on the probability of smoking cessation over time using the model-based or naïve standard errors with the ‘best’ working correlation from Q2(f)?

```r
# m3.toep$geese$vbeta.naiv
# #Check that the Wald Statistic Computed using the Robust Estimation Matrix same as in summary(fit.ar1
# (coef(m3.toep)/sqrt(diag(m3.toep$geese$vbeta)))^2
# #obtain p-values based on Robust aka Empirical Estimates
# round(pchisq((coef(m3.toep)/sqrt(diag(m3.toep$geese$vbeta)))^2,1,lower.tail=F),4)
# #Compute the Wald Statistic using the Naive aka Model-based Estimation Matrix
# (coef(m3.toep)/sqrt(diag(m3.toep$geese$vbeta.naiv)))^2
# #obtain p-values from Naive aka Model-Based Estiamtes
# round(pchisq((coef(m3.toep)/sqrt(diag(m3.toep$geese$vbeta.naiv)))^2,1,lower.tail=F),4)
```

```r
library(aod)
sigma = m3.toep$geese$vbeta.naiv
b = coef(m3.toep)
wald.test(sigma,b,Terms = c(8,9),df=2,verbose = FALSE)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 6.0, df = 2, P(> X2) = 0.049
##
## F test:
## W = 3.0, df1 = 2, df2 = 2, P(> W) = 0.25
```

The conclusion p = 0.049 < 0.05 therefore we reject the null hypothesis and conclude that the interaction of White race and time are significant in our full marginal model with the ‘best’ working correlation structure found in (e).

(iii) 2 points: comment on the magnitude of the standard errors from Q2(d)(i), Q2(g)(i) and Q2(g)(ii)

```
#Q2(d)(i)
sqrt(diag((m1.ind$geese$vbeta)))[8:9]
```

```
## [1] 0.04645 0.00178
```

```
#Q2(g)(i)
sqrt(diag((m3.toep$geese$vbeta)))[8:9]
```

```
## [1] 0.04582 0.00173
```

```
#Q2(g)(ii)
sqrt(diag((m3.toep$geese$vbeta.naiv)))[8:9]
```

```
## [1] 0.04531 0.00176
```

All of them are very similar.

(iv) 5 points: Following the instructions for Q2(d)(ii), what is your inference for the effect of Group on the probability of smoking cessation over time using the sandwich estimator with the 'best' working correlation from Q2(f)?

H_0: Coefficient of Socsup * Month = Coefficient of Ctrl * Month = Coefficient of Discussion * Month = Coefficient of Socsup * Month2 = Coefficient of Ctrl * Month2 = Coefficient of Discussion * Month2 = 0

$$H_0 : \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0, H_1 : \text{At least one of betas is nonzero.}$$

```
m3.toep.test2 <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + rac

anova(m3.toep,m3.toep.test2)
```

```
## Analysis of 'Wald statistic' Table
##
## Model 1 quit ~ control + discussion + socsup + racew + month + month2 + racew * month + racew * mont
## Model 2 quit ~ control + discussion + socsup + racew + month + month2 + racew * month + racew * mont
##   Df   X2 P(>|Chi|)
## 1  6 8.38      0.21
```

The conclusion p = 0.21 > 0.05 therefore we fail to reject the null hypothesis and conclude that the interaction of time and group are not significant.

(v) 5 points: Similarly, what is your inference for the effect of Group on the probability of smoking cessation over time using the model-based or naïve standard errors with the 'best' working correlation from Q2(f)?

```
library(aod)
sigma = m3.toep$geese$vbeta.naiv
b = coef(m3.toep)
wald.test(sigma,b,Terms = c(10:15),df=2,verbose = FALSE)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 9.1, df = 6, P(> X2) = 0.17
```

```
##
## F test:
## W = 1.5, df1 = 6, df2 = 2, P(> W) = 0.45
```

The conclusion p = 0.17 > 0.05 therefore we fail to reject the null hypothesis and conclude that the interaction of time and group are not significant.

(vi) 2 points: comment on the magnitude of the standard errors from Q2(d)(ii), Q2(g)(iv) and Q2(g)(v). Why are they similar or not similar?

```
#Q2(d)(ii)
sqrt(diag((m1.ind$geese$vbeta)))[10:15]
```

```
## [1] 0.05624 0.00212 0.06285 0.00241 0.05509 0.00210
```

```
#Q2(g)(iv)
sqrt(diag((m3.toep$geese$vbeta)))[10:15]
```

```
## [1] 0.05484 0.00205 0.06132 0.00233 0.05433 0.00206
```

```
#Q2(g)(v)
sqrt(diag((m3.toep$geese$vbeta.naiv)))[10:15]
```

```
## [1] 0.05662 0.00219 0.05687 0.00224 0.04944 0.00194
```

They are quite similar. It might be due to the fact that the estimates are not impacted by the correlation structure, as well as the result given by sandwich estimator is similar to the model-based one.

(h) 2 points: based on the final model in Q2(d) and the final models in Q2(g) (the final model based on the sandwich estimator with the 'best' working correlation from Q2(f), and the final model based on the 'model-based' or naïve standard errors using the 'best' working correlation from Q2(f)), what is your observation for the magnitude of the parameter estimates for the fixed effects across the different final models? Why are they similar or not similar?

In Q2(d), white race and time are not significant, and the interaction of time and group are not significant. In Q2(g),
- Using sandwich estimator (part (i) & (iv)): the interaction of White race and time are not significant, and the interaction of time and group are not significant.
- Using model-based or naïve standard errors using the 'best' working correlation((ii)& (v)): the interaction of White race and time are significant & the interaction of time and group are not significant.

```
# q2d model
q2d.model <- geeglm(quit ~ control + discussion + socsup +racew + month + month2 , data=datasmoking, fa

# q2f model using sandwich estimator
q2f_sandwich <-  geeglm(quit ~ control + discussion + socsup +racew + month + month2, family = binomial

# q2f model using model-based or naive SE
q2f_naive <-  geeglm(quit ~ control + discussion + socsup +racew + month + month2 + racew*month + racew

summary(q2d.model)
```

```
##
## Call:
## geeglm(formula = quit ~ control + discussion + socsup + racew +
##     month + month2, family = binomial(link = "logit"), data = datasmoking,
##     id = id, corstr = "independence", std.err = "san.se")
```

```
##
##  Coefficients:
##              Estimate   Std.err  Wald Pr(>|W|)
## (Intercept) -1.079424  0.152906 49.84  1.7e-12 ***
## control     -0.393076  0.244573  2.58   0.1080
## discussion   0.191537  0.237942  0.65   0.4208
## socsup       0.581535  0.204133  8.12   0.0044 **
## racew        0.427464  0.200831  4.53   0.0333 *
## month       -0.112441  0.021084 28.44  9.7e-08 ***
## month2       0.003940  0.000806 23.88  1.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.01     0.1
## Number of clusters:   489  Maximum cluster size: 4
```

```
summary(q2f_sandwich)
```

```
##
## Call:
## geeglm(formula = quit ~ control + discussion + socsup + racew +
##     month + month2, family = binomial(link = "logit"), data = datasmoking,
##     id = id, zcor = zcor.toep, corstr = "userdefined", std.err = "san.se")
##
##  Coefficients:
##             Estimate  Std.err  Wald Pr(>|W|)
## (Intercept) -1.05451  0.15452 46.57  8.8e-12 ***
## control     -0.41814  0.24065  3.02   0.0823 .
## discussion   0.23044  0.23305  0.98   0.3228
## socsup       0.63378  0.20044 10.00   0.0016 **
## racew        0.35083  0.19894  3.11   0.0778 .
## month       -0.11669  0.02079 31.49  2.0e-08 ***
## month2       0.00411  0.00079 27.02  2.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = userdefined
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.01   0.106
##   Link = identity
##
## Estimated Correlation Parameters:
##         Estimate Std.err
## alpha:1    0.409  0.0589
## alpha:2    0.294  0.0563
## alpha:3    0.240  0.0622
## Number of clusters:   489  Maximum cluster size: 4
```

```
summary(q2f_naive)
```

```
##
## Call:
## geeglm(formula = quit ~ control + discussion + socsup + racew +
##     month + month2 + racew * month + racew * month2, family = binomial(link = "logit"),
##     data = datasmoking, id = id, zcor = zcor.toep, corstr = "userdefined",
##     std.err = "san.se")
##
##  Coefficients:
##               Estimate   Std.err  Wald Pr(>|W|)
## (Intercept)  -0.950061  0.157979 36.17  1.8e-09 ***
## control      -0.445879  0.240098  3.45   0.0633 .
## discussion    0.227619  0.234355  0.94   0.3314
## socsup        0.634339  0.201233  9.94   0.0016 **
## racew        -0.052382  0.240373  0.05   0.8275
## month        -0.151633  0.024841 37.26  1.0e-09 ***
## month2        0.005265  0.000946 30.96  2.6e-08 ***
## racew:month   0.130173  0.045324  8.25   0.0041 **
## racew:month2 -0.004255  0.001739  5.98   0.0144 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = userdefined
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.02   0.125
##   Link = identity
##
## Estimated Correlation Parameters:
##         Estimate Std.err
## alpha:1    0.410  0.0652
## alpha:2    0.295  0.0603
## alpha:3    0.244  0.0649
## Number of clusters:   489  Maximum cluster size: 4
```

They are quite similar. It might be due to the fact that the estimates are not impacted by the correlation structure.

(i) 4 points: Using the estimates from the final model in Q2(d), what is the estimated OR for social support group vs. control for smoking cessation at time 6 months? What is it when you use the final model from Q2(g) with the model-based or naïve standard errors using the 'best' working correlation structure in Q2(f)?

```
m1_pred <- unique(data.frame(socsup=datasmoking$socsup,control=datasmoking$control,discussion=datasmokin
m1_pred_num <- m1_pred[m1_pred$month == 6  & m1_pred$socsup == 1,]
m1_pred_denom <- m1_pred[m1_pred$month == 6  & m1_pred$control == 1,]
exp(c(m1_pred_num$lp - m1_pred_denom$lp))
```

```
## [1] 1.73 4.06
```

```
m2_pred <- unique(data.frame(socsup=datasmoking$socsup,control=datasmoking$control,discussion=datasmokin
m2_pred_num <- m2_pred[m2_pred$month == 6  & m2_pred$socsup == 1,]
```

```
m2_pred_denom <- m2_pred[m2_pred$month == 6  & m2_pred$control == 1,]
exp(c(m2_pred_num$lp - m2_pred_denom$lp))
```

```
## [1] 1.66 5.24
```

Using the estimates from the final model in Q2(d), the estimated OR for social support group vs. control for smoking cessation at time 6 months is 4.06 for white group, and 1.73 for non-white group.
Using the estimates from the final model in Q2(f), the estimated OR for social support group vs. control for smoking cessation at time 6 months is 5.24 for white group, and 1.65 for non-white group.