SurvMeth 687 FA23 HW #2

Cheng, Chia Wen

2023-10-24

Setup

```r
# set working direction to where the data set was stored locally
setwd("G:/My Drive/0. study abroad/academic/10. 2023 Fall/5. SurvMeth 687 Applications of Statistical Mo
# load required packages
library(lme4)
library(nlme)
library(lmerTest)
library(tidyverse)
library(dplyr)
library(haven) # for read_dta()
library(ggplot2)
library(RColorBrewer)
library(geepack)
# read the data set from STATA data
growth <- read_dta('growth.dta')
# looking into data structure
str(growth) # subject is numeric; gender is character; age is numeric; distance is numeric; index is nu
```

```
## tibble [108 x 6] (S3: tbl_df/tbl/data.frame)
##  $ subject  : num [1:108] 1 1 1 1 2 2 2 2 3 3 ...
##   ..- attr(*, "label")= chr "subject"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ gender   : chr [1:108] "F" "F" "F" "F" ...
##   ..- attr(*, "label")= chr "gender"
##   ..- attr(*, "format.stata")= chr "%1s"
##  $ age      : num [1:108] 8 10 12 14 8 10 12 14 8 10 ...
##   ..- attr(*, "label")= chr "age in years"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ distance : num [1:108] 21 20 21.5 23 21 21.5 24 25.5 20.5 24 ...
##   ..- attr(*, "label")= chr "distance (mm) from center of pituitary to pteryo-maxillary fissure"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ index    : num [1:108] 1 1 1 1 2 2 2 2 3 3 ...
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ largedist: num [1:108] 0 0 0 0 0 0 1 1 0 1 ...
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  - attr(*, "label")= chr "growth dataset written by Stat/Transfer Ver. 16.3.1592.0705"
```

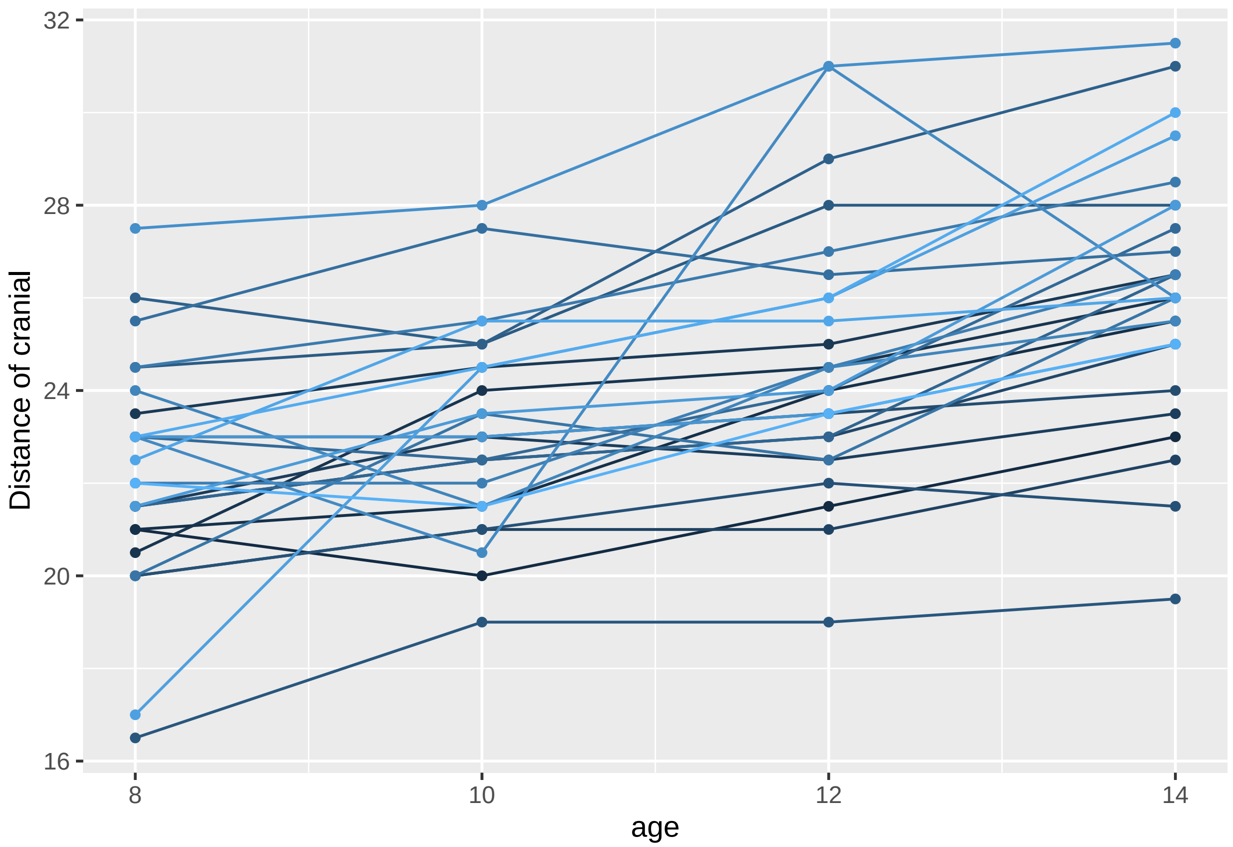
```
typeof(growth) # dataset growth is a list
```

```
## [1] "list"
```

```
# examine whether any columns contain NAs
apply(is.na(growth), 2, sum) # no NA’s
```

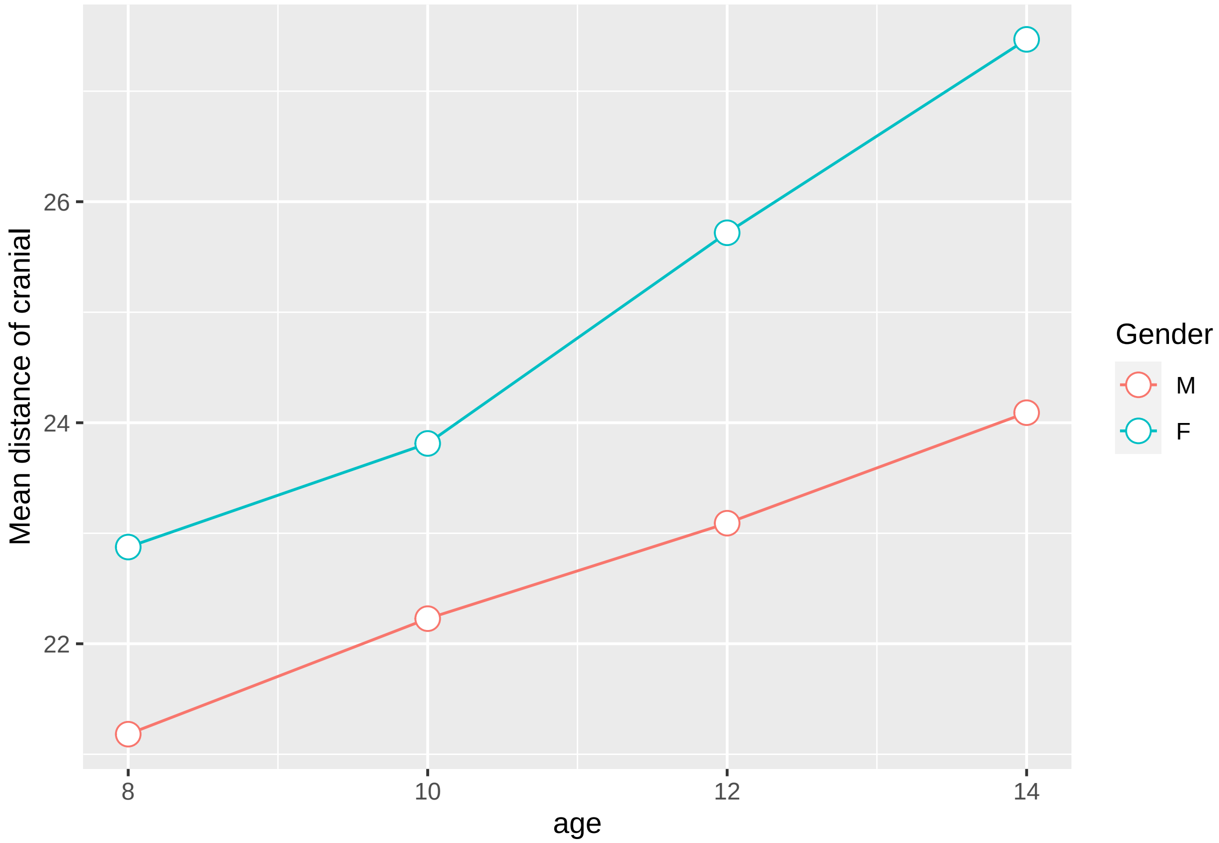
```
##   subject    gender       age  distance     index largedist
##         0         0         0         0         0         0
```

```
# spaghetti plot for distance of each subject: plotting each subject's cranial distance
ggplot(data = growth, aes(x = age, y = distance, group = subject, colour = subject)) +
  geom_line() +
  geom_point(size = 2, shape = 20) +
  ylab("Distance of cranial") +
  theme(legend.position = "none")
```



```
# create an aggregate data set
aggdata <- aggregate(growth$distance,
                     by = list(growth$age, growth$gender),
                     FUN = mean, na.rm = TRUE)
colnames(aggdata) <- c("age", "gender", "mean_distance")
# spaghetti plot for mean distances: indicating the relationships between age and distance grouped by g
```

```r
ggplot(data = aggdata, aes(x = age, y = mean_distance,
                           group = factor(gender),
                           colour = factor(gender, labels = c("M", "F")))) +
  geom_line() + geom_point(size = 4, shape = 21, fill = "white") +
  ylab("Mean distance of cranial") +
  labs(color = "Gender")
```



According to the last homework, there is no NA value in our dataset. I first generated a spaghetti plot to see the distribution of the data by each subject. The plot shows that slopes and intercepts vary from subject to subject while some individuals seem to have relatively stable growth trajectories during age 8-14.
I then created an aggregated data to more easily make a spaghetti plot with a focus on the effects of ages and genders instead of subjects (because we're not fitting multilevel models). The plot indicates that the expected trajectory of cranial growth across ages we collected may be different because of genders: female shows a relatively constant growth trajectory across ages, while male shows distinct differences in cranial growth trajectory between ages 8-10 and ages 10-14. In addition, male seems to have similar expected trajectory of cranial growth as female from age 8 to age 10.

Please perform the following analyses to address the three study objectives, using your judgment on how to keep things as simple as possible. You will need to provide clearly annotated code that was used for these analyses with your homework submission.

1. Fit several marginal models to the DISTANCE dependent variable using maximum likelihood estimation methods, in each case choosing a different variance-covariance structure for the residuals within each individual child. Include all of the fixed effects necessary to address the first two research objectives in each of the models; for now, treat AGE as a continuous predictor. Which of the models appears to have the best fit? Provide a table of relevant fit criteria to support this decision.

```r
# center the intercept to make an intercept at 0 meaningful
growth[, 7] <- growth$age - 8
colnames(growth)[7] <- "centered_age"
# mutate a new column named "year" to present consecutive numbers for ages
growth[, 8] <- growth$age
colnames(growth)[8] <- "year"
growth["year"][growth["year"] == 8] <- 1
growth["year"][growth["year"] == 10] <- 2
growth["year"][growth["year"] == 12] <- 3
growth["year"][growth["year"] == 14] <- 4

# fit the first marginal model: use unstructured variance-covariance matrix for errors
mar_mod_1 <- gls(distance ~ centered_age + gender + centered_age*gender,
                 na.action = "na.omit",
                 correlation = corSymm(form = ~ year | subject),
                 # unique values of the covariate for "corSymm" objects
                 # must be a sequence of consecutive integers
                 weights = varIdent(form = ~ 1 | year),
                 method = "REML",
                 data = growth)
summary(mar_mod_1)
```

```
## Generalized least squares fit by REML
##   Model: distance ~ centered_age + gender + centered_age * gender
##   Data: growth
##        AIC      BIC    logLik
##   452.5468 489.5683 -212.2734
##
## Correlation Structure: General
##  Formula: ~year | subject
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.568
## 3 0.659 0.581
## 4 0.522 0.725 0.740
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | year
##  Parameter estimates:
##         1         2         3         4
## 1.0000000 0.8788793 1.0744596 0.9586879
##
## Coefficients:
##                          Value Std.Error  t-value p-value
## (Intercept)          21.236286 0.6128510 34.65163  0.0000
## centered_age          0.476365 0.0991583  4.80408  0.0000
## genderM               1.220426 0.7961168  1.53297  0.1283
## centered_age:genderM  0.350439 0.1288104  2.72058  0.0076
##
##  Correlation:
##                      (Intr) cntrd_ gendrM
## centered_age         -0.381
## genderM              -0.770  0.293
## centered_age:genderM  0.293 -0.770 -0.381
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.34272823 -0.63481502 -0.07904148  0.63772263  2.16523296
##
## Residual standard error: 2.329213
## Degrees of freedom: 108 total; 104 residual
```

```
getVarCov(mar_mod_1)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 5.4252 2.7092 3.8411 2.7152
## [2,] 2.7092 4.1906 2.9745 3.3137
## [3,] 3.8411 2.9745 6.2632 4.1333
## [4,] 2.7152 3.3137 4.1333 4.9862
##   Standard Deviations: 2.3292 2.0471 2.5026 2.233
```

```
anova(mar_mod_1, type = "marginal", adjustSigma = F)
```

```
## Denom. DF: 104
##                      numDF   F-value p-value
## (Intercept)              1 1200.7354  <.0001
## centered_age             1   23.0792  <.0001
## gender                   1    2.3500  0.1283
## centered_age:gender      1    7.4016  0.0076
```

```
# fit the second marginal model: use an autocorrelation structure of order 1 variance-covariance matrix
mar_mod_2 <- gls(distance ~ centered_age + gender + centered_age*gender,
                 na.action = "na.omit",
                 correlation = corAR1(form = ~ year | subject),
                 weights = varIdent(form = ~ 1 | year),
                 method = "REML",
                 data = growth)
summary(mar_mod_2)
```

```
## Generalized least squares fit by REML
##   Model: distance ~ centered_age + gender + centered_age * gender
##   Data: growth
##        AIC      BIC    logLik
##   460.7962 484.5957 -221.3981
##
## Correlation Structure: AR(1)
##  Formula: ~year | subject
##  Parameter estimate(s):
##       Phi
## 0.6332579
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | year
##  Parameter estimates:
##         1         2         3         4 
## 1.0000000 0.8875549 1.0172458 0.8833068 
## 
## Coefficients:
##                          Value Std.Error   t-value p-value
## (Intercept)          21.213895 0.7009285 30.265418  0.0000
## centered_age          0.482699 0.1390171  3.472228  0.0008
## genderM               1.426336 0.9105329  1.566485  0.1203
## centered_age:genderM  0.300851 0.1805885  1.665947  0.0987
## 
##  Correlation: 
##                      (Intr) cntrd_ gendrM
## centered_age         -0.672              
## genderM              -0.770  0.517       
## centered_age:genderM  0.517 -0.770 -0.672
## 
## Standardized residuals:
##         Min          Q1         Med          Q3         Max 
## -2.33898020 -0.62982298 -0.07384351  0.57391277  2.13028356 
## 
## Residual standard error: 2.411406 
## Degrees of freedom: 108 total; 104 residual
```

```
getVarCov(mar_mod_2)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 5.8149 3.2683 2.3721 1.3044
## [2,] 3.2683 4.5807 3.3246 1.8281
## [3,] 2.3721 3.3246 6.0172 3.3087
## [4,] 1.3044 1.8281 3.3087 4.5369
##   Standard Deviations: 2.4114 2.1403 2.453 2.13
```

```
anova(mar_mod_2, type = "marginal", adjustSigma = F)
```

```
## Denom. DF: 104 
##                     numDF  F-value p-value
## (Intercept)             1 915.9955  <.0001
## centered_age            1  12.0564  0.0008
## gender                  1   2.4539  0.1203
## centered_age:gender     1   2.7754  0.0987
```

```
# fit the third marginal model: use an exponential spatial correlation structure of variance-covariance
mar_mod_3 <- gls(distance ~ centered_age + gender + centered_age*gender,
                 na.action = "na.omit",
                 correlation = corCompSymm(form = ~ year | subject),
                 method = "REML",
                 data = growth)
summary(mar_mod_3)
```

```
## Generalized least squares fit by REML
##   Model: distance ~ centered_age + gender + centered_age * gender
##   Data: growth
##        AIC      BIC    logLik
##   445.7572 461.6236 -216.8786
##
## Correlation Structure: Compound symmetry
##  Formula: ~year | subject
##  Parameter estimate(s):
##       Rho
## 0.6318381
##
## Coefficients:
##                          Value Std.Error  t-value p-value
## (Intercept)          21.209091 0.6497598 32.64143  0.0000
## centered_age          0.479545 0.0934699  5.13048  0.0000
## genderM               1.406534 0.8440627  1.66639  0.0986
## centered_age:genderM  0.304830 0.1214209  2.51052  0.0136
##
##  Correlation:
##                      (Intr) cntrd_ gendrM
## centered_age         -0.432
## genderM              -0.770  0.332
## centered_age:genderM  0.332 -0.770 -0.432
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.45773173 -0.57853118 -0.07360637  0.58204364  2.29634478
##
## Residual standard error: 2.284881
## Degrees of freedom: 108 total; 104 residual
```

```
getVarCov(mar_mod_3)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 5.2207 3.2986 3.2986 3.2986
## [2,] 3.2986 5.2207 3.2986 3.2986
## [3,] 3.2986 3.2986 5.2207 3.2986
## [4,] 3.2986 3.2986 3.2986 5.2207
##   Standard Deviations: 2.2849 2.2849 2.2849 2.2849
```

```
anova(mar_mod_3, type = "marginal", adjustSigma = F)
```

```
## Denom. DF: 104
##                      numDF   F-value p-value
## (Intercept)              1 1065.4633  <.0001
## centered_age             1   26.3218  <.0001
## gender                   1    2.7768  0.0986
## centered_age:gender      1    6.3027  0.0136
```

```
# a table of relevant fit criteria--AIC, BIC, and Log Likelihood (can't use LRT because the models are
cus_tab_1 <- data.frame("AIC" = summary(mar_mod_1)$"AIC", "BIC" = summary(mar_mod_1)$"BIC", "logLik" =
cus_tab_2 <- data.frame("AIC" = summary(mar_mod_2)$"AIC", "BIC" = summary(mar_mod_2)$"BIC", "logLik" = s
cus_tab_3 <- data.frame("AIC" = summary(mar_mod_3)$"AIC", "BIC" = summary(mar_mod_3)$"BIC", "logLik" = s
cus_tab <- rbind(cus_tab_1, cus_tab_2, cus_tab_3)
cus_tab
```

```
##        AIC      BIC    logLik
## 1 452.5468 489.5683 -212.2734
## 2 460.7962 484.5957 -221.3981
## 3 445.7572 461.6236 -216.8786
```

Since our dependent variable of interest, `distance`, is a continuous, numeric variable, I use `gls()` function to fit marginal models.
The third marginal model with compound symmetry variance-covariance structure appears to have the best fit among the three marginal models because it has the smallest AIC and BIC at 453 and 490 respectively.

2. Interpret the estimated fixed effects and each of the estimated variances and covariances for the residuals in the best fitting model from Part 1. Do your overall inferences change when comparing the estimates to those generated from the multilevel modeling approach in Homework 1? What can you not make inference about in this marginal model?

```
mar_mod_3$coefficients
```

```
##          (Intercept)         centered_age              genderM
##           21.2090909            0.4795455            1.4065341
## centered_age:genderM
##            0.3048295
```

```
anova(mar_mod_3, type = "marginal", adjustSigma = F)
```

```
## Denom. DF: 104
##                      numDF   F-value p-value
## (Intercept)              1 1065.4633  <.0001
## centered_age             1   26.3218  <.0001
## gender                   1    2.7768  0.0986
## centered_age:gender      1    6.3027  0.0136
```

```
getVarCov(mar_mod_3)
```

```
## Marginal variance covariance matrix
##        [,1]   [,2]   [,3]   [,4]
## [1,] 5.2207 3.2986 3.2986 3.2986
## [2,] 3.2986 5.2207 3.2986 3.2986
## [3,] 3.2986 3.2986 5.2207 3.2986
## [4,] 3.2986 3.2986 3.2986 5.2207
##   Standard Deviations: 2.2849 2.2849 2.2849 2.2849
```

According to the results of the third marginal model I fit with an compound symmetry variance-covariance structure:
- Intercept: Across our sample, holding all other predictors constant, the mean cranial distance for female subjects at age 8 is 21.209 mm. This is statistically significant at 0.05 level and thus we cannot refuse the null

hypothesis. We can conclude that the mean cranial distance for 8-year-old female subjects is significantly different from 0.
- centered_age: Across our sample, holding all other predictors constant, each one-year increase in age is expected to be associated with a 0.48-mm growth in cranial distance. This is statistically significant at 0.05 level and thus we cannot refuse the null hypothesis. We can conclude that the expected annual cranial distance increase is significantly different from 0.
- genderM: Across our sample, holding all other predictors constant, the mean cranial distance for male subjects at age 8 is expected to be 1.407 mm larger than 8-year-old female subjects. This is not statistically significant at 0.05 level and thus we can refuse the null hypothesis. We can conclude that the expected difference between 8-year-old male's and 8-year-old female's mean cranial distances may not be significantly different from 0.
- centered_age:genderM: Across our sample, holding all other predictors constant, each one-year increase in age is expected to be associated with an additional 0.305-mm increase for male subjects compared to female subjects. This is statistically significant at 0.05 level and thus we cannot refuse the null hypothesis. We can conclude that the expected annual cranial distance increase of male subjects is significantly different from the expected annual cranial distance increase of female subjects.
- variance and covariance for the residuals: Since I selected a compound symmetry structure, the correlations of residuals are presumed to be the same for each pair of estimators. The variance of residuals for each estimator is 5.22, and the correlation of residuals for each pair of predictors is 3.30.
- The interpretations of fixed effects in this homework did not change compared to Homework 1. However, since I did not have regression models focusing on differences between each subject, I am not able to make inferences about the between-group difference. I can only discuss about the overall trends, where each subject is expected to share the same pattern for each estimator.

3. Refit the best-fitting model from Part 1 using GEE, with an appropriate working correlation matrix based on your earlier findings (you only need to fit one model, with a working correlation matrix that approximates that best variance-covariance matrix found earlier). Do any of your inferences about the fixed effects change, compared to Part 2?

```
# fit GEE model with exchangeable variance-covariance structure
gee_mod_1 <- geeglm(distance ~ centered_age + gender + centered_age*gender,
                    id = subject, waves = age,
                    family = gaussian("identity"),
                    data = growth,
                    corstr = "exchangeable")
# geepack makes the anova() function available, which is nice for overall tests of factors / terms in t
summary(gee_mod_1)
```

```
##
## Call:
## geeglm(formula = distance ~ centered_age + gender + centered_age *
##     gender, family = gaussian("identity"), data = growth, id = subject,
##     waves = age, corstr = "exchangeable")
##
##  Coefficients:
##                      Estimate  Std.err     Wald Pr(>|W|)
## (Intercept)          21.20909  0.56043 1432.185  < 2e-16 ***
## centered_age          0.47955  0.06313   57.697 3.05e-14 ***
## genderM               1.40653  0.77380    3.304   0.0691 .
## centered_age:genderM  0.30483  0.11687    6.803   0.0091 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    4.905   1.015
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.6178  0.1312
## Number of clusters:   27  Maximum cluster size: 4
```

```
anova(gee_mod_1)
```

```
## Analysis of 'Wald statistic' Table
## Model: gaussian, link: identity
## Response: distance
## Terms added sequentially (first to last)
##
##                     Df   X2 P(>|Chi|)
## centered_age         1 89.1    <2e-16 ***
## gender               1  9.6    0.0020 **
## centered_age:gender  1  6.8    0.0091 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interpretation for the fixed effects in this question had slightly changed compared to question two because the standard deviations changed and resulted in changes in p-values. Based on the ANOVA, all estimates remain the same but "gender" is now statistically significant at 0.05 level.

4. Perform some descriptive analyses of the binary LARGEDIST variable, looking at estimated probabilities of having larger than expected growth as a function of AGE and GENDER. You can create tables and/or figures for this analysis. Discuss the descriptive results.

```
# create a contingency table with counts for largedist by age and gender
contingency_tab <- table(growth$age, growth$gender, growth$largedist)
contingency_tab
```

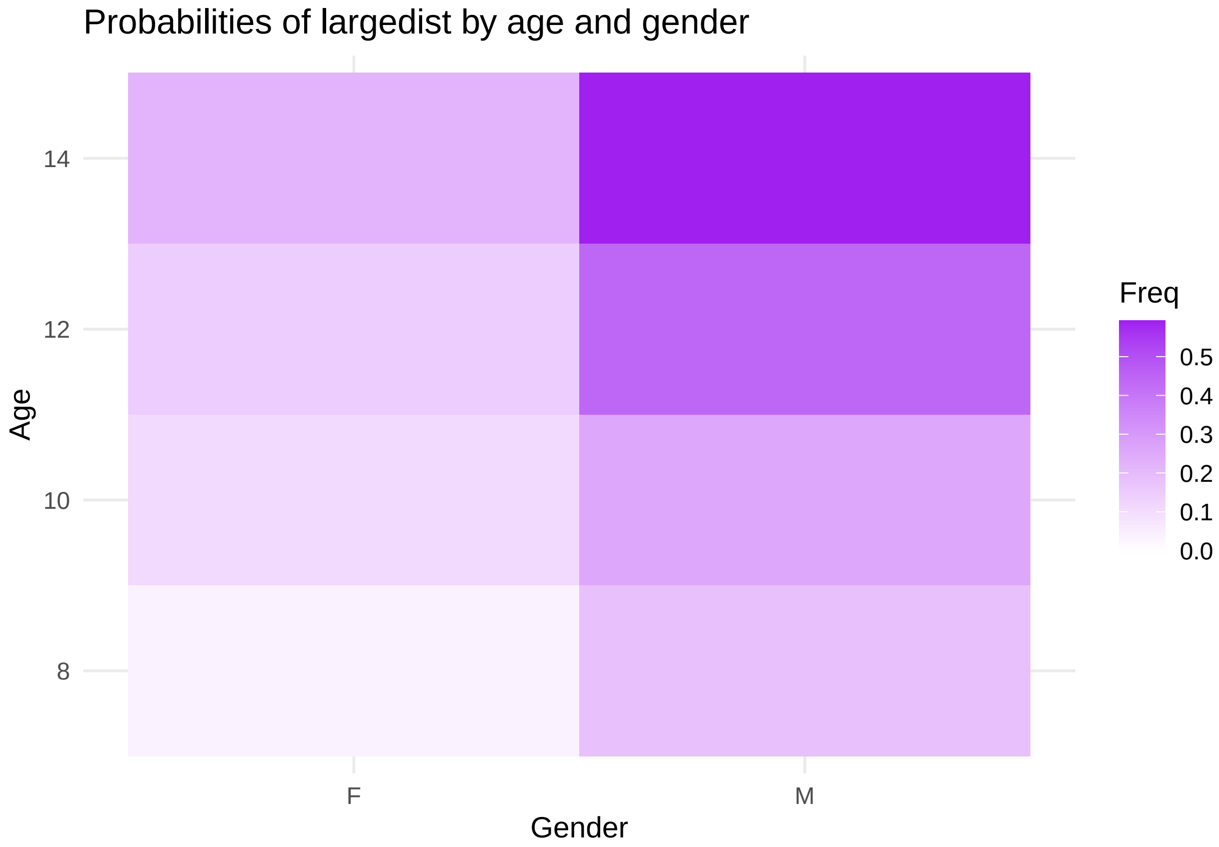
```
## , ,  = 0
##
##
##       F  M
##   8  10 11
##   10  8  9
##   12  7  4
##   14  5  0
##
## , ,  = 1
##
##
##       F  M
##   8   1  5
##   10  3  7
##   12  4 12
##   14  6 16
```

```
# create a probability table
probability_tab <- contingency_tab / rowSums(contingency_tab)
probability_tab
```

```
## , ,  = 0
##
##
##             F       M
##   8  0.37037 0.40741
##   10 0.29630 0.33333
##   12 0.25926 0.14815
##   14 0.18519 0.00000
##
## , ,  = 1
##
##
##             F       M
##   8  0.03704 0.18519
##   10 0.11111 0.25926
##   12 0.14815 0.44444
##   14 0.22222 0.59259
```

```
# create a heatmap of the probabilities
heatmap_data <- as.data.frame(probability_tab)

ggplot(heatmap_data, aes(x = Var2, y = Var1, fill = Freq)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "purple") +
  labs(title = "Probabilities of largedist by age and gender", x = "Gender", y = "Age") +
  theme_minimal()
```



I first created contingency tables for the binary variable `largedist` to see the counts of genders and ages within each value of `largedist`. I then created probability tables based on the contingency tables, and plotted a heatmap based on the probabilities.
From the heatmap, at each age level, male subjects constantly have higher probabilities to have biennial cranial growth larger than expected compared to female subjects. However, both female subjects and male subjects have increased probabilities of having biennial cranial growth larger than expected compared to the latest age stage.

5. Fit several marginal logit models to the binary LARGEDIST dependent variable using GEE. Each model should have the same fixed effects necessary to fully test the interaction between GENDER and AGE (continuous) when predicting this binary outcome, but a different working correlation matrix. Which model would seem to have the best fit, and why?

```
# fit the first GEE logit model, using independent variance-covariance structure
gee_mod_bi_1 <- geeglm(largedist ~ centered_age + gender + centered_age*gender,
                       id = subject, waves = age,
                       family = binomial("logit"),
                       data = growth,
                       corstr = "independence")
summary(gee_mod_bi_1)
```

```
##
## Call:
## geeglm(formula = largedist ~ centered_age + gender + centered_age *
##     gender, family = binomial("logit"), data = growth, id = subject,
##     waves = age, corstr = "independence")
```

```
##
##  Coefficients:
##                       Estimate Std.err Wald Pr(>|W|)
## (Intercept)             -1.985   0.781 6.47   0.0110 *
## centered_age             0.368   0.133 7.68   0.0056 **
## genderM                  0.808   0.978 0.68   0.4085
## centered_age:genderM     0.275   0.178 2.40   0.1216
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation structure = independence
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.921   0.166
## Number of clusters:   27  Maximum cluster size: 4
```

```
anova(gee_mod_bi_1)
```

```
## Analysis of ’Wald statistic’ Table
## Model: binomial, link: logit
## Response: largedist
## Terms added sequentially (first to last)
##
##                     Df    X2 P(>|Chi|)
## centered_age         1 24.09   9.2e-07 ***
## gender               1  4.88     0.027 *
## centered_age:gender  1  2.40     0.122
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
QIC(gee_mod_bi_1)
```

```
##       QIC      QICu Quasi Lik       CIC    params      QICC
##    125.03    120.61    -56.31      6.21      4.00    126.85
```

```
# fit the second GEE logit model, using exchangeable variance-covariance structure
gee_mod_bi_2 <- geeglm(largedist ~ centered_age + gender + centered_age*gender,
                       id = subject, waves = age,
                       family = binomial("logit"),
                       data = growth,
                       corstr = "exchangeable")
summary(gee_mod_bi_2)
```

```
##
## Call:
## geeglm(formula = largedist ~ centered_age + gender + centered_age *
##     gender, family = binomial("logit"), data = growth, id = subject,
##     waves = age, corstr = "exchangeable")
##
##  Coefficients:
##                       Estimate Std.err Wald Pr(>|W|)
## (Intercept)            -1.962   0.777 6.37   0.0116 *
## centered_age            0.366   0.131 7.79   0.0053 **
## genderM                 0.634   0.999 0.40   0.5258
## centered_age:genderM    0.260   0.180 2.10   0.1477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.915   0.204
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.405   0.137
## Number of clusters:   27  Maximum cluster size: 4
```

```r
anova(gee_mod_bi_2)
```

```
## Analysis of 'Wald statistic' Table
## Model: binomial, link: logit
## Response: largedist
## Terms added sequentially (first to last)
##
##                      Df    X2 P(>|Chi|)
## centered_age          1 24.04   9.5e-07 ***
## gender                1  4.69      0.03 *
## centered_age:gender  1  2.10      0.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
QIC(gee_mod_bi_2)
```

```
##       QIC      QICu Quasi Lik       CIC    params      QICC
##    126.18    121.00    -56.50      6.59      4.00    129.04
```

```r
# fit the third GEE logit model, using AR(1) variance-covariance structure
gee_mod_bi_3 <- geeglm(largedist ~ centered_age + gender + centered_age*gender,
                       id = subject, waves = age,
                       family = binomial("logit"),
                       data = growth,
                       corstr = "ar1")
summary(gee_mod_bi_3)
```

```
##
## Call:
## geeglm(formula = largedist ~ centered_age + gender + centered_age *
##     gender, family = binomial("logit"), data = growth, id = subject,
##     waves = age, corstr = "ar1")
##
```

```
##  Coefficients:
##                      Estimate Std.err Wald Pr(>|W|)
## (Intercept)            -2.030   0.797 6.48   0.0109 *
## centered_age            0.374   0.137 7.50   0.0062 **
## genderM                 0.883   0.990 0.80   0.3725
## centered_age:genderM    0.228   0.183 1.56   0.2119
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.909   0.186
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.552   0.134
## Number of clusters:   27  Maximum cluster size: 4
```

```r
anova(gee_mod_bi_3)
```

```
## Analysis of ’Wald statistic’ Table
## Model: binomial, link: logit
## Response: largedist
## Terms added sequentially (first to last)
##
##                      Df    X2 P(>|Chi|)
## centered_age          1 23.63   1.2e-06 ***
## gender                1  5.31     0.021 *
## centered_age:gender   1  1.56     0.212
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
QIC(gee_mod_bi_3)
```

```
##       QIC      QICu Quasi Lik       CIC    params      QICC
##     125.1     120.7     -56.4       6.2       4.0     128.0
```

```r
# fit the fourth GEE logit model, using unstructured variance-covariance structure
gee_mod_bi_4 <- geeglm(largedist ~ centered_age + gender + centered_age*gender,
                       id = subject, waves = age,
                       family = binomial("logit"),
                       data = growth,
                       corstr = "unstructured")
summary(gee_mod_bi_4)
```

```
##
## Call:
## geeglm(formula = largedist ~ centered_age + gender + centered_age *
##     gender, family = binomial("logit"), data = growth, id = subject,
##     waves = age, corstr = "unstructured")
##
##  Coefficients:
##                       Estimate Std.err Wald Pr(>|W|)  
## (Intercept)             -1.932   0.775 6.21    0.013 *
## centered_age             0.344   0.134 6.59    0.010 *
## genderM                  0.852   0.960 0.79    0.375  
## centered_age:genderM     0.274   0.193 2.01    0.156  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation structure = unstructured 
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     0.91   0.162
##   Link = identity 
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.530   0.237
## alpha.1:3    0.444   0.178
## alpha.1:4    0.134   0.113
## alpha.2:3    0.763   0.247
## alpha.2:4    0.233   0.140
## alpha.3:4    0.286   0.129
## Number of clusters:   27  Maximum cluster size: 4
```

```r
anova(gee_mod_bi_4)
```

```
## Analysis of ’Wald statistic’ Table
## Model: binomial, link: logit
## Response: largedist
## Terms added sequentially (first to last)
##
##                      Df    X2 P(>|Chi|)    
## centered_age          1 23.87     1e-06 ***
## gender                1  7.21    0.0072 ** 
## centered_age:gender   1  2.01    0.1560    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
QIC(gee_mod_bi_4)
```

```
##       QIC      QICu Quasi Lik       CIC    params      QICC 
##    124.45    120.68    -56.34      5.88      4.00    138.20
```

```r
# create a table to store all relative fit criteria for comparison
qic_tab <- data.frame("gee model binary 1" = QIC(gee_mod_bi_1), "gee model binary 2" = QIC(gee_mod_bi_2
qic_tab
```

```
##           gee.model.binary.1 gee.model.binary.2 gee.model.binary.3
## QIC                   125.03             126.18              125.1
## QICu                  120.61             121.00              120.7
## Quasi Lik             -56.31             -56.50              -56.4
## CIC                     6.21               6.59                6.2
## params                  4.00               4.00                4.0
## QICC                  126.85             129.04              128.0
##          gee.model.binary.4
## QIC                   124.45
## QICu                  120.68
## Quasi Lik             -56.34
## CIC                     5.88
## params                  4.00
## QICC                  138.20
```

The fourth model with unstructured variance-covariance structure seems to have the best fit because it has the smallest QIC among the four models.

6. Interpret the estimated fixed effects in the best-fitting GEE model from Part 5.

```
exp(gee_mod_bi_4$coefficients)
```

```
##          (Intercept)         centered_age              genderM
##                0.145                1.411                2.344
## centered_age:genderM
##                1.316
```

```
anova(gee_mod_bi_4)
```

```
## Analysis of ’Wald statistic’ Table
## Model: binomial, link: logit
## Response: largedist
## Terms added sequentially (first to last)
##
##                     Df    X2 P(>|Chi|)
## centered_age         1 23.87     1e-06 ***
## gender               1  7.21    0.0072 **
## centered_age:gender  1  2.01    0.1560
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

According to the results of the fourth GEE logit model I fit with an unstructured variance-covariance structure:
- Intercept: Across our sample, holding all other predictors constant, the odds of an 8-year-old female subject having their cranial growth larger than expected is 0.145.
- centered_age: Across our sample, holding all other predictors constant, the odds of a subject having their cranial growth larger than expected is 1.411 times higher when age increases by one unit. This is statistically significant at 0.05 level and thus we cannot refuse the null hypothesis. We can conclude that the odds of having larger annual cranial growth than expected is significantly different from 0.
- genderM: Across our sample, holding all other predictors constant, the odds of an 8-year-old male subject having their cranial growth larger than expected is 2.344 times higher than an 8-year-old female subject. This is statistically significant at 0.05 level and thus we cannot refuse the null hypothesis. We can conclude that the difference of odds between 8-year-old male and 8-year-old female subjects in having larger cranial

growth than expected is significantly different from 0.
- centered_age:genderM: Across our sample, holding all other predictors constant, the odds of male subjects having their cranial growth larger than expected is 1.316 times higher than female subjects when age increases by one unit. This is not statistically significant at 0.05 level and thus we can refuse the null hypothesis. We can conclude that the odds of having their annual cranial growth larger than expected is of no difference between male and female subjects.

7. Provide a one-page writeup, summarizing the methods used in your analyses and the conclusions that you would draw based on the final results for each of your best-fitting models. Keep the original research questions in mind when discussing your conclusions.

To study the expected trajectory of cranial growth across ages we collected, the effect of gender on the expected trajectory of cranial growth, and whether there is significant differences between the genders on the probability of having larger than expected growth, I fitted several models including marginal models with three variance-covariance structure using `gls()` and generalized estimating equations (GEE) models with four variance-covariance structure using `geeglm()`.

The results indicate that:
1. From the marginal models:
* The best-fitting model is the one with compound symmetry variance-covariance structure because it has the smallest AIC and BIC values among the three models.
* Across our sample, holding all other predictors constant, the mean cranial distance for 8-year-old female subjects is 21.209 mm and this is significantly different from 0.
* Across our sample, holding all other predictors constant, each one-year increase in age is expected to be associated with a 0.48-mm growth in cranial distance and the expected annual growth is also significantly different from 0.
* Across our sample, holding all other predictors constant, the mean cranial distance for male subjects at age 8 is expected to be 1.407 mm larger than 8-year-old female subjects but the difference in expected cranial distance between 8-year-old males and same age female is not significant.
* Across our sample, holding all other predictors constant, each one-year increase in age is expected to be associated with an additional 0.305-mm increase for male subjects compared to female subjects and this difference is significant from 0.
* With a compound symmetry structure, the correlations of residuals are presumed to be the same, where the variance of residuals for each estimator is 5.22, and the correlation of residuals for each pair of predictors is 3.30.
* If changed to fit a GEE model instead of a GLS marginal model, all estimates remain the same but “gender” becomes statistically significant.

2. For GEE logit models:
* The best-fitting model is the one with unstructured variance-covariance structure because it has the smallest QIC value among the four models.
* Across our sample, holding all other predictors constant, the odds of an 8-year-old female subject having their cranial growth larger than expected is 0.145.
* Across our sample, holding all other predictors constant, the odds of a subject having their cranial growth larger than expected is 1.411 times higher when age increases by one unit and this is highly significant.
* Across our sample, holding all other predictors constant, the odds of an 8-year-old male subject having their cranial growth larger than expected is 2.344 times higher than an 8-year-old female subject and this is highly significant.
* Across our sample, holding all other predictors constant, the odds of male subjects having their cranial growth larger than expected is 1.316 times higher than female subjects when age increases by one unit but the odds difference between male and female subjects for every one-year increase is not significant.

To conclude, the expected trajectory of annual cranial growth is not 0 and gender plays a role in differentiating each year’s expected increase, but gender does not enlarge the difference at age 8. For both genders, the probability of having larger than expected growth increases as subjects grow older. In addition, 8-year-old male subjects have significantly higher probabilities to have larger cranial distance compared to 8-year-old female subjects. However, gender does not play a role in differentiating the probabilities of having larger than expected annual cranial growth.