BIS 628 HW 2

Joanna Chen

2/4/2020

Data Description

Background for the data set:
In a study of dental growth, measurements of the distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure were obtained on 11 girls and 16 boys at ages 8, 10, 12, and 14 (Potthoff and Roy, 1964).
Dataset: dental-data.txt
Each row of the data contains the following six variables (wide format):
ID Gender Yi8 Yi10 Yi12 Yi14
ID: patient ID number
Gender: 'M' or 'F'
Yij: dental growth (mm) at 8 years old (j=8), 10 years old (j=10), 12 years old (j=12), and 14 years old (j=14)

Question 1

(a) Read 'dental-data.txt' file into SAS or R (depending upon your preference) and put the data in a 'long' format, with 4 rows per subject.

```
library(ggplot2)
library(dplyr)
library(data.table)
library(nlme)

setwd("~/Downloads/BIS628HW2")
dt = read.csv("dental-data-1.txt", header = F, sep = "")
colnames(dt) = c("ID","gender","Y_i8","Y_i10","Y_i12","Y_i14")

dt.long = reshape2::melt(dt, id.vars = c('ID','gender')) # reshape the data
setDT(dt.long)
setnames(dt.long,"value","growth")

dt.long$year = substring(dt.long$variable, 4) # extract month
dt.long[,variable:=NULL] # delete the variable column since we already extract month from it
dt.long$cyear = factor(dt.long$year, levels = c(8,10,12,14)) # create cmonth variable which treats mont
dt.long$year = as.numeric(dt.long$year)
knitr::kable(head(dt.long[order(ID)]))
```

ID	gender	growth	year	cyear
1	F	21.0	8	8
1	F	20.0	10	10
1	F	21.5	12	12
1	F	23.0	14	14
2	F	21.0	8	8
2	F	21.5	10	10

(b) On a single graph, construct a time plot that displays mean growth (mm) versus age (in years) for boys

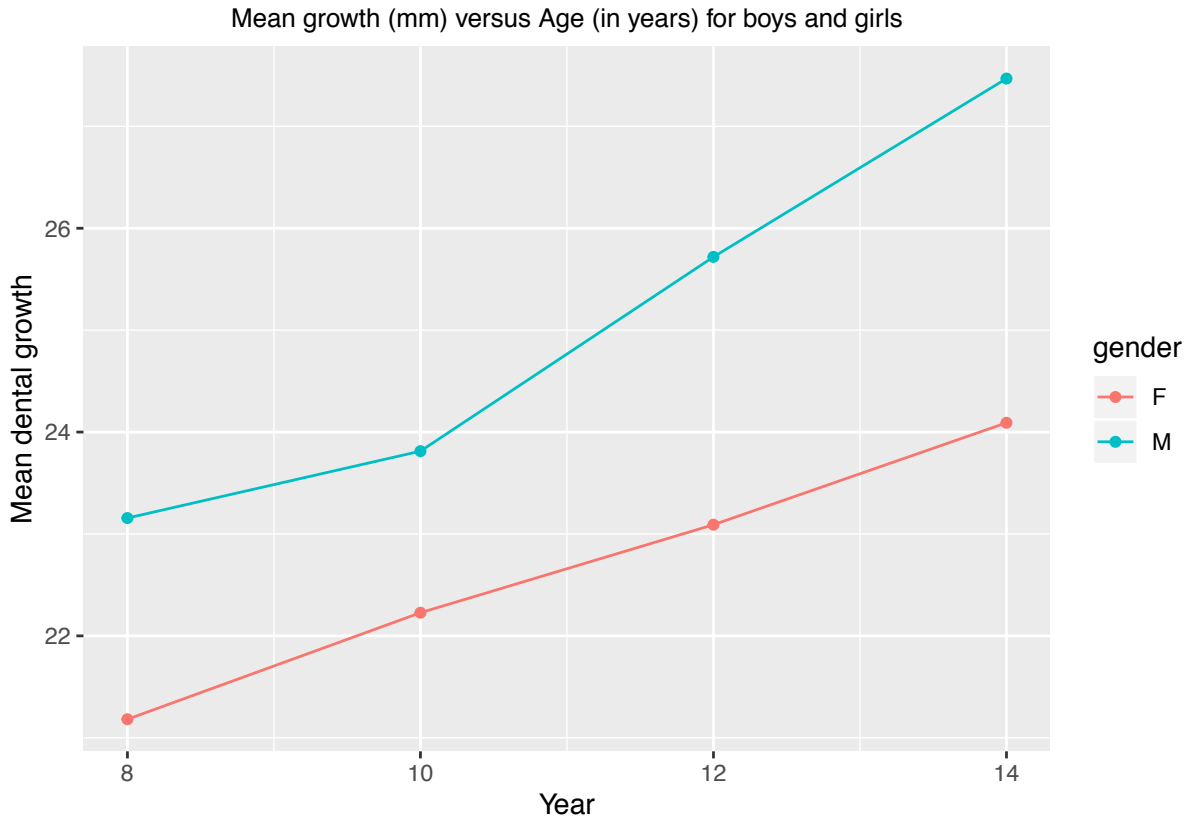


Figure 1: Mean growth (mm) versus Age (in years) for boys and girls.

and girls. Briefly describe the time trends for boys and girls.

```
mean = dt.long %>%
  group_by(gender, year) %>%
  summarize(mean_growth = mean(growth))
# mean

p1 = ggplot(mean, aes(x = year, y = mean_growth, color = gender, group = gender )) +
  geom_point() +
  geom_line() +
  labs(title="Mean growth (mm) versus Age (in years) for boys and girls",
       x="Year", y="Mean dental growth", color = "gender") +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
p1
```

Both genders' dental growth mean are increasing from year 8 to year 14. Female group's growth approximatly linear, while male group's growth looks piecewise linear with a knot at time 10 and their growth becomes more quicker since Year 10 becuase the slope increases.

Question 2

Consider the following covariance structures (Ri): Unstructured, Compound Symmetry, Autoregressive, Toeplitz, Exponential, and Heterogeneous Compound Symmetry. Identify All nested covariance models:
(a) Write out Ri for each covariance model using 4 time points
(1) Unstructured:

$$R_i(Unstructured) : \text{Cov}\,(e_i) = \text{Cov}\,(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}$$

Note that $\sigma_{jk} = \sigma_{kj}$ for j = 1,2,3,4 and k =1,2,3,4.

(2) Compound Symmetry:

$$R_i(CS) : \text{Cov}\,(e_i) = \text{Cov}\,(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

(3) First Order Autoregressive (AR1):

$$R_i(AR1) : \text{Cov}\,(e_i) = \text{Cov}\,(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

(4) Toeplitz:

$$R_i(Toeplitz) : \text{Cov}\,(e_i) = \text{Cov}\,(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$$

(5) Exponential:
We knoe that

$$R_i : \text{Cov}\,(e_{ij}, e_{ik}) = \text{Cov}\,(Y_{ij}, Y_{ik}) = \sigma^2 \exp\left(-\theta \left|\mathbf{t}_{ij} - \mathbf{t}_{ik}\right|\right), \text{ where } \theta = -\log(\rho).$$

Therefore,

$$R_i(Exponential) : \sigma^2 \begin{pmatrix} 1 & e^{-2\theta} & e^{-4\theta} & e^{-6\theta} \\ e^{-2\theta} & 1 & e^{-2\theta} & e^{-4\theta} \\ e^{-4\theta} & e^{-2\theta} & 1 & e^{-2\theta} \\ e^{-6\theta} & e^{-4\theta} & e^{-2\theta} & 1 \end{pmatrix}$$

(6) Heterogeneous Compound Symmetry

$$R_i(HeterCS) : \text{Cov}\,(e_i) = \text{Cov}\,(Y_i) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \rho^3\sigma_1\sigma_4 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho^2\sigma_2\sigma_4 \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho^3\sigma_1\sigma_4 & \rho^2\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2 \end{pmatrix}$$

(b) Show what constraints you can apply to an Ri to get nested models and identify All nested models. (For example, what constraints would you apply to the Unstructured covariance to get to the Compound Symmetry covariance?)

(1) Nested within Unstuctured:

- CS is nested within Unstuctured if $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$ and σ_{ij} where i $\neq$ j (or off diagnol terms) $= \rho\sigma^2$.

- AR1 is nested within Unstuctured and they are equivalent if $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$ and
  $\sigma_{12} = \sigma_{21} = \sigma_{23} = \sigma_{32} = \sigma_{34} = \sigma_{43} = \sigma^2\rho$ and
  $\sigma_{13} = \sigma_{31} = \sigma_{24} = \sigma_{42} = \sigma^2\rho^2$ and
  $\sigma_{14} = \sigma_{41} = \sigma^2\rho^3$
- Toeplitz is nested within Unstuctured if $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$ and
  $\sigma_{12} = \sigma_{21} = \sigma_{23} = \sigma_{32} = \sigma_{34} = \sigma_{43} = \sigma^2\rho_1$ and
  $\sigma_{13} = \sigma_{31} = \sigma_{24} = \sigma_{42} = \sigma^2\rho_2$ and
  $\sigma_{14} = \sigma_{41} = \sigma^2\rho_3$
- Exponential is nested within Unstuctured and they are equivalent if $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$ and
  $\sigma_{12} = \sigma_{21} = \sigma_{23} = \sigma_{32} = \sigma_{34} = \sigma_{43} = \sigma^2 e^{-2\theta}$ and
  $\sigma_{13} = \sigma_{31} = \sigma_{24} = \sigma_{42} = \sigma^2 e^{-4\theta}$ and
  $\sigma_{14} = \sigma_{41} = \sigma^2 e^{-6\theta}$
- Heter CS is nested within Unstuctured if for the off-diagnol terms, for i $\neq$ j, $\sigma_{ij} = \sigma_i\sigma_j\rho^{|i-j|}$.

(2) Nested within Toeplitz

- CS is nested within Toeplitz and they are equivalent if $\rho_1 = \rho_2 = \rho_3 = \rho$
- AR1 is nested within Toeplitz if $\rho_1 = \rho, \rho_2 = \rho^2, \rho_3 = \rho^3$
- Exp is nested within Toeplitz if $\rho_1 = e^{-2\theta}, \rho_2 = e^{-4\theta}, \rho_3 = e^{-6\theta}$,

(3) Nested within Exponential and Nested within AR1

- AR1 and Exponential are equivalent if we let $\rho = e^{-2\theta}$. Moreover, we talked about exponential can be reduced to AR1. It makes sense here because of our balanced design.

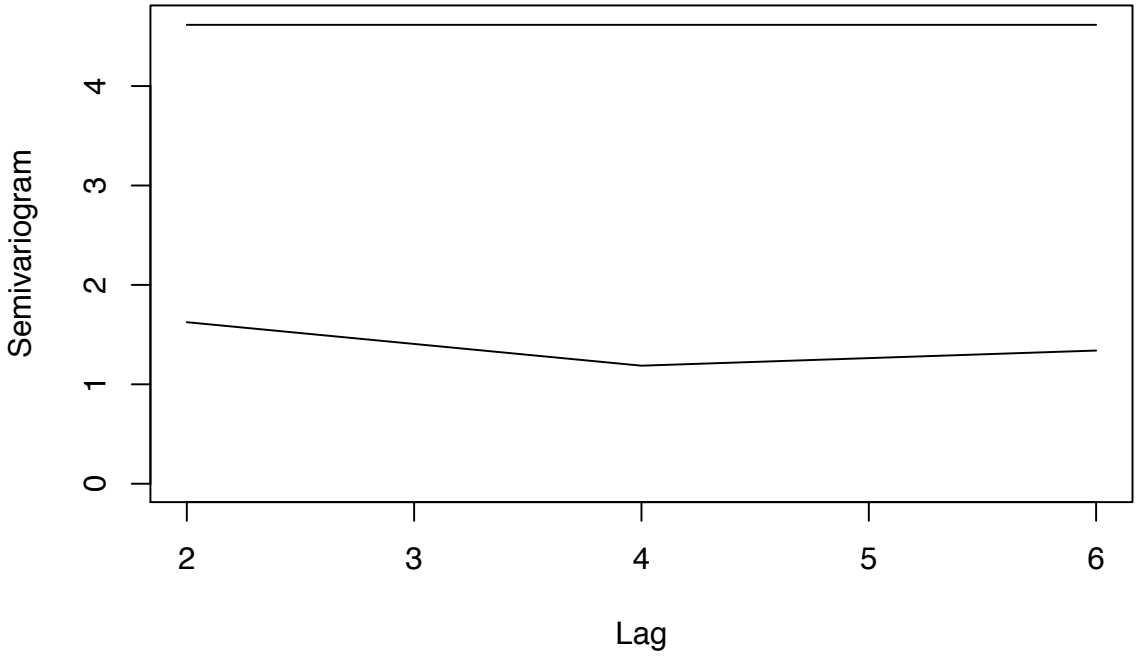
(4) Nested within HCS

- CS is nested within HCS if $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma$. They are equivalent when $\sigma = 1$.

Question 3

(a) Using the "maximal" model for the mean response, plot estimated empirical semivariogram for the dataset.

```
source("variogram.R")
#library(lmenssp) #Sample variogram
#Fit a piecewise linear model with independent covariance structure
lm <- lm(growth ~ gender*cyear, data=dt.long)
#Plot sample variogram
variogram(resid=resid(lm),timeVar=dt.long$year,id=dt.long$ID,irregular=F)
```




```
## $lags
## [1] 2 4 6
##
## $means
##        2        4        6
## 1.624947 1.187513 1.339752
##
## $sizes
##  2  4  6
## 81 54 27
##
## $process.var
## [1] 4.61609
```

(b) Describe the different sources of variability in the response (variance) using the plot of the empirical semivariogram.
At lag 2, the semivariogram is close to 2, therefore we can conclude that there are measurement error variance close to 2. The variance of the serial process is a little decreasing with the time increasing and became flat since lag 4, which shows that the correlation increases a little and keep unchanged. Finally, since the seal is close to 5 , the rest are variance of the random effect (between subjects variability).
We can see that the variance of the serial process is little and almost no change. That means ρ is almost no change. It implicitly shows that CS is the best in this case.

Question 4 (30 points):

For the "maximal" model for the mean response, assume a saturated model for the mean response.

(a) Fit the following models for the covariance:

Independent (R matrix structure)

```r
#Add occasion index for each subject
dt.long = dt.long %>% group_by(ID) %>% mutate(time = row_number())

m1 <- gls(growth ~ cyear*gender, data=dt.long)
summary(m1) #Get estimates of the fixed effects and correlation matrix
```

```
## Generalized least squares fit by REML
##   Model: growth ~ cyear * gender
##   Data: dt.long
##        AIC      BIC    logLik
##   481.8314 505.2779 -231.9157
##
## Coefficients:
##                     Value Std.Error  t-value p-value
## (Intercept)     21.181818 0.6688878 31.66722  0.0000
## cyear10          1.045455 0.9459502  1.10519  0.2717
## cyear12          1.909091 0.9459502  2.01817  0.0463
## cyear14          2.909091 0.9459502  3.07531  0.0027
## genderM          1.974432 0.8689108  2.27231  0.0252
## cyear10:genderM -0.389205 1.2288254 -0.31673  0.7521
## cyear12:genderM  0.653409 1.2288254  0.53173  0.5961
## cyear14:genderM  1.403409 1.2288254  1.14207  0.2562
##
##  Correlation:
##                 (Intr) cyer10 cyer12 cyer14 gendrM cy10:M cy12:M
## cyear10         -0.707
## cyear12         -0.707  0.500
## cyear14         -0.707  0.500  0.500
## genderM         -0.770  0.544  0.544  0.544
## cyear10:genderM  0.544 -0.770 -0.385 -0.385 -0.707
## cyear12:genderM  0.544 -0.385 -0.770 -0.385 -0.707  0.500
## cyear14:genderM  0.544 -0.385 -0.385 -0.770 -0.707  0.500  0.500
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.11040070 -0.66206138 -0.08195731  0.61307909  2.38060370
##
## Residual standard error: 2.21845
## Degrees of freedom: 108 total; 100 residual
```

Unstructured covariance

```r
m2 <- gls(growth ~ cyear * gender,
          data = dt.long,
          corr = corSymm(form = ~ time | ID),
          weights = varIdent(form = ~ 1 | cyear)
)
summary(m2)
```

```
## Generalized least squares fit by REML
##   Model: growth ~ cyear * gender
##   Data: dt.long
##        AIC     BIC    logLik
##   430.8239 477.717 -197.4119
```

```
##
## Correlation Structure: General
##  Formula: ~time | ID
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.689
## 3 0.774 0.563
## 4 0.684 0.726 0.728
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | cyear
##  Parameter estimates:
##        8       10       12       14
## 1.000000 1.015268 1.261007 1.108182
##
## Coefficients:
##                     Value Std.Error  t-value p-value
## (Intercept)     21.181818 0.6075172 34.86621  0.0000
## cyear10          1.045455 0.4827438  2.16565  0.0327
## cyear12          1.909091 0.4855948  3.93145  0.0002
## cyear14          2.909091 0.5129109  5.67173  0.0000
## genderM          1.974432 0.7891879  2.50185  0.0140
## cyear10:genderM -0.389205 0.6271026 -0.62064  0.5362
## cyear12:genderM  0.653409 0.6308061  1.03583  0.3028
## cyear14:genderM  1.403409 0.6662909  2.10630  0.0377
##
##  Correlation:
##                 (Intr) cyer10 cyer12 cyer14 gendrM cy10:M cy12:M
## cyear10         -0.378
## cyear12         -0.031  0.072
## cyear14         -0.287  0.536  0.421
## genderM         -0.770  0.291  0.023  0.221
## cyear10:genderM  0.291 -0.770 -0.055 -0.413 -0.378
## cyear12:genderM  0.023 -0.055 -0.770 -0.324 -0.031  0.072
## cyear14:genderM  0.221 -0.413 -0.324 -0.770 -0.287  0.536  0.421
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.32359081 -0.65778202 -0.08816554  0.59313651  2.15580725
##
## Residual standard error: 2.014906
## Degrees of freedom: 108 total; 100 residual
```

Compound symmetry covariance

```
m3 <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID))
summary(m3)
```

```
## Generalized least squares fit by REML
##   Model: growth ~ cyear * gender
##   Data: dt.long
```

```
##       AIC      BIC   logLik
##   424.932 450.9837 -202.466
##
## Correlation Structure: Compound symmetry
##  Formula: ~time | ID
##  Parameter estimate(s):
##       Rho
## 0.6858426
##
## Coefficients:
##                     Value Std.Error  t-value p-value
## (Intercept)     21.181818 0.6688878 31.66722  0.0000
## cyear10          1.045455 0.5302026  1.97180  0.0514
## cyear12          1.909091 0.5302026  3.60068  0.0005
## cyear14          2.909091 0.5302026  5.48675  0.0000
## genderM          1.974432 0.8689107  2.27231  0.0252
## cyear10:genderM -0.389205 0.6887534 -0.56509  0.5733
## cyear12:genderM  0.653409 0.6887534  0.94868  0.3451
## cyear14:genderM  1.403409 0.6887534  2.03761  0.0442
##
##  Correlation:
##                 (Intr) cyer10 cyer12 cyer14 gendrM cy10:M cy12:M
## cyear10         -0.396
## cyear12         -0.396  0.500
## cyear14         -0.396  0.500  0.500
## genderM         -0.770  0.305  0.305  0.305
## cyear10:genderM  0.305 -0.770 -0.385 -0.385 -0.396
## cyear12:genderM  0.305 -0.385 -0.770 -0.385 -0.396  0.500
## cyear14:genderM  0.305 -0.385 -0.385 -0.770 -0.396  0.500  0.500
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.11040071 -0.66206139 -0.08195731  0.61307909  2.38060371
##
## Residual standard error: 2.21845
## Degrees of freedom: 108 total; 100 residual
```

First-order autoregressive covariance

```
m4 <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corAR1(form= ~ time | ID))
summary(m4)
```

```
## Generalized least squares fit by REML
##   Model: growth ~ cyear * gender
##   Data: dt.long
##       AIC      BIC    logLik
##   442.1506 468.2023 -211.0753
##
## Correlation Structure: AR(1)
##  Formula: ~time | ID
##  Parameter estimate(s):
##       Phi
```

```
## 0.6402953
##
## Coefficients:
##                     Value Std.Error  t-value p-value
## (Intercept)     21.181818 0.6618643 32.00326  0.0000
## cyear10          1.045455 0.5613801  1.86229  0.0655
## cyear12          1.909091 0.7189820  2.65527  0.0092
## cyear14          2.909091 0.8038276  3.61905  0.0005
## genderM          1.974432 0.8597870  2.29642  0.0237
## cyear10:genderM -0.389205 0.7292541 -0.53370  0.5947
## cyear12:genderM  0.653409 0.9339850  0.69959  0.4858
## cyear14:genderM  1.403409 1.0442027  1.34400  0.1820
##
##  Correlation:
##                 (Intr) cyer10 cyer12 cyer14 gendrM cy10:M cy12:M
## cyear10         -0.424
## cyear12         -0.543  0.640
## cyear14         -0.607  0.492  0.734
## genderM         -0.770  0.326  0.418  0.467
## cyear10:genderM  0.326 -0.770 -0.493 -0.379 -0.424
## cyear12:genderM  0.418 -0.493 -0.770 -0.565 -0.543  0.640
## cyear14:genderM  0.467 -0.379 -0.565 -0.770 -0.607  0.492  0.734
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.13279559 -0.66908697 -0.08282701  0.61958488  2.40586590
##
## Residual standard error: 2.195156
## Degrees of freedom: 108 total; 100 residual
```

Toeplitz

```
m5 <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corARMA(form= ~ time | ID, p=3, q=0)) #4 times point here, then p = 4-1=3.
summary(m5)
```

```
## Generalized least squares fit by REML
##   Model: growth ~ cyear * gender
##   Data: dt.long
##        AIC      BIC    logLik
##   426.1938 457.4559 -201.0969
##
## Correlation Structure: ARMA(3,0)
##  Formula: ~time | ID
##  Parameter estimate(s):
##        Phi1        Phi2        Phi3
##  0.31868858  0.56194935 -0.03633786
##
## Coefficients:
##                     Value Std.Error   t-value p-value
## (Intercept)     21.181818 0.6700137 31.614007  0.0000
## cyear10          1.045455 0.5481706  1.907170  0.0594
## cyear12          1.909091 0.4739588  4.027968  0.0001
```

```
## cyear14          2.909091 0.6166385  4.717660  0.0000
## genderM          1.974432 0.8703733  2.268488  0.0254
## cyear10:genderM -0.389205 0.7120945 -0.546563  0.5859
## cyear12:genderM  0.653409 0.6156906  1.061262  0.2911
## cyear14:genderM  1.403409 0.8010370  1.751990  0.0828
##
##  Correlation:
##                 (Intr) cyer10 cyer12 cyer14 gendrM cy10:M cy12:M
## cyear10         -0.409
## cyear12         -0.354  0.432
## cyear14         -0.460  0.675  0.521
## genderM         -0.770  0.315  0.272  0.354
## cyear10:genderM  0.315 -0.770 -0.333 -0.519 -0.409
## cyear12:genderM  0.272 -0.333 -0.770 -0.401 -0.354  0.432
## cyear14:genderM  0.354 -0.519 -0.401 -0.770 -0.460  0.675  0.521
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.10685435 -0.66094885 -0.08181959  0.61204886  2.37660330
##
## Residual standard error: 2.222184
## Degrees of freedom: 108 total; 100 residual
```

Exponential

```
m6 <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corExp(form= ~ time | ID))
summary(m6)
```

```
## Generalized least squares fit by REML
##   Model: growth ~ cyear * gender
##   Data: dt.long
##        AIC      BIC    logLik
##   442.1506 468.2023 -211.0753
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~time | ID
##  Parameter estimate(s):
##    range
## 2.243029
##
## Coefficients:
##                     Value Std.Error  t-value p-value
## (Intercept)     21.181818 0.6618643 32.00326  0.0000
## cyear10          1.045455 0.5613801  1.86229  0.0655
## cyear12          1.909091 0.7189820  2.65527  0.0092
## cyear14          2.909091 0.8038276  3.61905  0.0005
## genderM          1.974432 0.8597870  2.29642  0.0237
## cyear10:genderM -0.389205 0.7292541 -0.53370  0.5947
## cyear12:genderM  0.653409 0.9339851  0.69959  0.4858
## cyear14:genderM  1.403409 1.0442027  1.34400  0.1820
##
##  Correlation:
```

```
##                 (Intr) cyer10 cyer12 cyer14 gendrM cy10:M cy12:M
## cyear10         -0.424
## cyear12         -0.543  0.640
## cyear14         -0.607  0.492  0.734
## genderM         -0.770  0.326  0.418  0.467
## cyear10:genderM  0.326 -0.770 -0.493 -0.379 -0.424
## cyear12:genderM  0.418 -0.493 -0.770 -0.565 -0.543  0.640
## cyear14:genderM  0.467 -0.379 -0.565 -0.770 -0.607  0.492  0.734
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.13279561 -0.66908697 -0.08282701  0.61958489  2.40586592
##
## Residual standard error: 2.195156
## Degrees of freedom: 108 total; 100 residual
```

Heterogeneous Compound Symmetry

```
m7 <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID),
          weights=varIdent(form= ~ 1 | cyear))
summary(m7)
```

```
## Generalized least squares fit by REML
##   Model: growth ~ cyear * gender
##   Data: dt.long
##        AIC      BIC    logLik
##   427.3399 461.2071 -200.6699
##
## Correlation Structure: Compound symmetry
##  Formula: ~time | ID
##  Parameter estimate(s):
##       Rho
## 0.6944064
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | cyear
##  Parameter estimates:
##        8       10       12       14
## 1.000000 1.046705 1.278928 1.110016
##
## Coefficients:
##                     Value Std.Error  t-value p-value
## (Intercept)     21.181818 0.6006204 35.26656  0.0000
## cyear10          1.045455 0.4812144  2.17253  0.0322
## cyear12          1.909091 0.5568193  3.42856  0.0009
## cyear14          2.909091 0.4991049  5.82862  0.0000
## genderM          1.974432 0.7802288  2.53058  0.0129
## cyear10:genderM -0.389205 0.6251159 -0.62261  0.5350
## cyear12:genderM  0.653409 0.7233294  0.90334  0.3685
## cyear14:genderM  1.403409 0.6483563  2.16456  0.0328
##
##  Correlation:
```

```
##                 (Intr) cyer10 cyer12 cyer14 gendrM cy10:M cy12:M
## cyear10         -0.341
## cyear12         -0.121  0.424
## cyear14         -0.276  0.464  0.424
## genderM         -0.770  0.262  0.093  0.212
## cyear10:genderM  0.262 -0.770 -0.326 -0.357 -0.341
## cyear12:genderM  0.093 -0.326 -0.770 -0.327 -0.121  0.424
## cyear14:genderM  0.212 -0.357 -0.327 -0.770 -0.276  0.464  0.424
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.3502718 -0.6642354 -0.0885678  0.5864574  2.1805617
##
## Residual standard error: 1.992033
## Degrees of freedom: 108 total; 100 residual
```

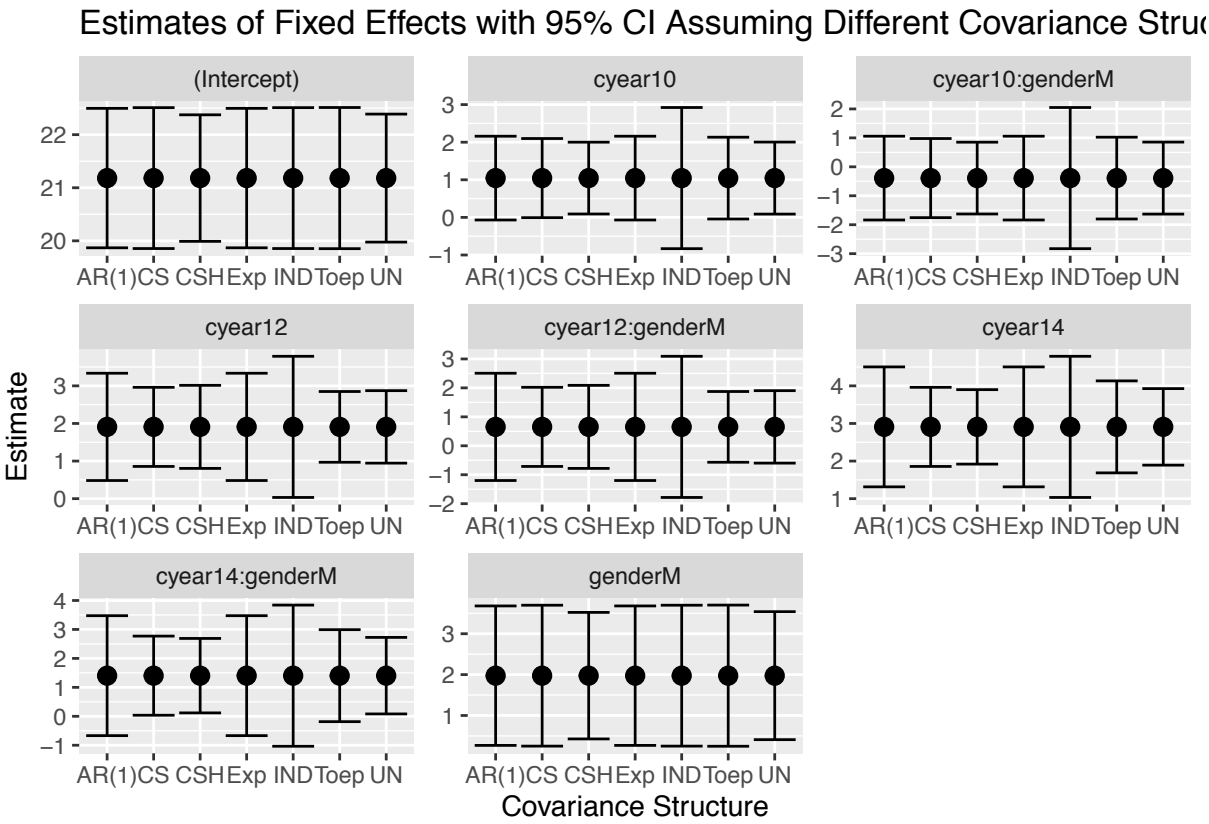
(b) From each covariance model in 3(a), plot parameter estimates and 95% CIs of the coefficients involved in testing the differences in the means of growth between boys and girls over time:

```r
#From Lab 3 code - function to extract estimates of fixed effects and 95% CIs
est <- function(model,cov) {
  a <- as.data.frame(intervals(model)$coef)
  a$covstr <- cov
  a$effect <- rownames(a)
  a
}

#Combine all the estimates of fixed effects and 95% CIs from different models
fixed <- rbind(est(m1,'IND'),est(m2,'UN'),est(m3,'CS'),est(m4,'AR(1)'),
               est(m5,'Toep') ,est(m6,'Exp'),est(m7,'CSH'))

#Plot estimates of fixed effects with 95% CI assuming different covariance structures (scatter plot wit
library(ggplot2)

p2 <- ggplot(fixed, aes(x=covstr,y=est.))
p2 + facet_wrap(~effect, scales = "free") + geom_point(size=3) +
  geom_errorbar(aes(ymin=lower,ymax=upper)) +
  labs(title="Estimates of Fixed Effects with 95% CI Assuming Different Covariance Structures",
       x="Covariance Structure", y="Estimate")
```



i. What do you notice? Comment on the width of 95%CIs.

The width of CI is different. The AR(1), Exp, IND model has wider CI which means less precise. (We can even tell IND model are not that good before testing.) We can also look at the interation terms that age10 and gender are not significant among all models. This is because the CI of all models applied on this interation term include 0 which means not significant.

ii. Comment on the statistical significance of the Type 3 Effects Test (Wald test or LRT) for the 'age*gender' parameters that test the gender difference in dental growth over time. For example, which covariance models contribute to the statistically significant results for the gender difference in dental growth over time ($p<0.05$)?

```
# LRT
m1_full <- gls(growth ~ cyear * gender, data=dt.long, method='ML')
m1_reduced <- gls(growth ~ cyear + gender, data=dt.long, method='ML')

m2_full <- gls(growth ~ cyear * gender,
          data = dt.long,
          corr = corSymm(form = ~ time | ID),
          weights = varIdent(form = ~ 1 | cyear),method='ML')
m2_reduced <- gls(growth ~ cyear + gender,
          data = dt.long,
          corr = corSymm(form = ~ time | ID),
          weights = varIdent(form = ~ 1 | cyear),method='ML')

m3_full <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID),method='ML')
m3_reduced <- gls(growth ~ cyear + gender,
```

```
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID),method='ML')

m4_full <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corAR1(form= ~ time | ID),method='ML')
m4_reduced <- gls(growth ~ cyear + gender,
          data=dt.long,
          corr=corAR1(form= ~ time | ID),method='ML')

m5_full <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corARMA(form= ~ time | ID, p=3, q=0),method='ML')
m5_reduced <- gls(growth ~ cyear + gender,
          data=dt.long,
          corr=corARMA(form= ~ time | ID, p=3, q=0),method='ML')

m6_full <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corExp(form= ~ time | ID),method='ML')
m6_reduced <- gls(growth ~ cyear + gender,
          data=dt.long,
          corr=corExp(form= ~ time | ID),method='ML')

m7_full <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID),
          weights=varIdent(form= ~ 1 | cyear),method='ML')
m7_reduced <- gls(growth ~ cyear + gender,
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID),
          weights=varIdent(form= ~ 1 | cyear),method='ML')


anova(m1_full, m1_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test L.Ratio p-value
## m1_full        1  9 488.2896 512.4288 -235.1448
## m1_reduced     2  6 484.9079 501.0007 -236.4539 1 vs 2 2.61827  0.4543
```

```
anova(m2_full, m2_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m2_full        1 18 431.7615 480.0399 -197.8808
## m2_reduced     2 15 433.9063 474.1382 -201.9531 1 vs 2 8.144741  0.0431
```

```
anova(m3_full, m3_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m3_full        1 10 426.6782 453.4995 -203.3391
## m3_reduced     2  7 428.7033 447.4782 -207.3517 1 vs 2 8.025105  0.0455
```

```
anova(m4_full, m4_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m4_full        1 10 445.2743 472.0956 -212.6372
```

```
## m4_reduced     2  7 443.2518 462.0267 -214.6259 1 vs 2 3.977508  0.2639
```

```
anova(m5_full, m5_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m5_full        1 12 427.7210 459.9066 -201.8605
## m5_reduced     2  9 430.3965 454.5356 -206.1982 1 vs 2 8.675414  0.0339
```

```
anova(m6_full, m6_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m6_full        1 10 445.2743 472.0956 -212.6372
## m6_reduced     2  7 443.2518 462.0267 -214.6259 1 vs 2 3.977508  0.2639
```

```
anova(m7_full, m7_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m7_full        1 13 428.7988 463.6665 -201.3994
## m7_reduced     2 10 431.1465 457.9678 -205.5732 1 vs 2 8.347683  0.0393
```

Therefore, Unstructured (m2), CS (m3), Toeplitz (m5), CSH (m7) contribute to the statistically significant results for the gender difference in dental growth over time ($p<0.05$) by LRT.

We also applied the Wald test here and got the same conclusion.

```
anova(m1,type = "marginal")
```

```
## Denom. DF: 100
##              numDF   F-value p-value
## (Intercept)      1 1002.8129  <.0001
## cyear            3    3.4307  0.0199
## gender           1    5.1634  0.0252
## cyear:gender     3    0.8180  0.4869
```

```
anova(m2,type = "marginal")
```

```
## Denom. DF: 100
##              numDF   F-value p-value
## (Intercept)      1 1215.6523  <.0001
## cyear            3   11.8554  <.0001
## gender           1    6.2593  0.0140
## cyear:gender     3    2.9341  0.0371
```

```
anova(m3,type = "marginal")
```

```
## Denom. DF: 100
##              numDF   F-value p-value
## (Intercept)      1 1002.8129  <.0001
## cyear            3   10.9205  <.0001
## gender           1    5.1634  0.0252
## cyear:gender     3    2.6037  0.0561
```

```
anova(m4,type = "marginal")
```

```
## Denom. DF: 100
##              numDF   F-value p-value
## (Intercept)      1 1024.2089  <.0001
## cyear            3    4.3695  0.0062
## gender           1    5.2735  0.0237
## cyear:gender     3    1.2530  0.2947
```

```r
anova(m5,type = "marginal")
```

```
## Denom. DF: 100
##              numDF  F-value p-value
## (Intercept)      1 999.4455  <.0001
## cyear            3   9.8562  <.0001
## gender           1   5.1460  0.0254
## cyear:gender     3   2.9277  0.0374
```

```r
anova(m6,type = "marginal")
```

```
## Denom. DF: 100
##              numDF   F-value p-value
## (Intercept)      1 1024.2089  <.0001
## cyear            3    4.3695  0.0062
## gender           1    5.2735  0.0237
## cyear:gender     3    1.2530  0.2947
```

```r
anova(m7,type = "marginal")
```

```
## Denom. DF: 100
##              numDF   F-value p-value
## (Intercept)      1 1243.7304  <.0001
## cyear            3   11.9893  <.0001
## gender           1    6.4038  0.0129
## cyear:gender     3    2.7792  0.0450
```

(c) Choose a model for the covariance that adequately fits the data:

i. Without conducting any statistical tests, are all covariance models considered in 3(a) applicable to the design of the study?

IND is not appropriate to use because we want to consider covariance. Other models are applicable because our design is balanced and equally-spaced.

ii. When choosing the best covariance based on statistical testing, show your work for selecting the best covariance for nested models

For nested model, we use LRT to compare. If results from LRT shows no difference, then we pick a reduced model.

```r
#Create a table of fit statistics - however, it doesn't give us p-value.
fitstats <- AIC(m1,m2,m3,m4,m5,m6,m7)
fitstats <- cbind(covstr=c('IND','UN','CS','AR(1)','Toep','Exp','CSH'), fitstats,
                  loglik=c(m1$logLik,m2$logLik,m3$logLik,m4$logLik,m5$logLik,m6$logLik,m7$logLik))
print(fitstats)
```

```
##    covstr df      AIC    loglik
## m1    IND  9 481.8314 -231.9157
## m2     UN 18 430.8239 -197.4119
## m3     CS 10 424.9320 -202.4660
## m4  AR(1) 10 442.1506 -211.0753
## m5   Toep 12 426.1938 -201.0969
## m6    Exp 10 442.1506 -211.0753
## m7    CSH 13 427.3399 -200.6699
```

Since every model is nested within UN, we do the following

```
anova(m2,m3)
```

```
##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m2     1 18 430.8239 477.7170 -197.4119
## m3     2 10 424.9320 450.9837 -202.4660 1 vs 2 10.10806  0.2575
```

```
anova(m2,m4)
```

```
##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m2     1 18 430.8239 477.7170 -197.4119
## m4     2 10 442.1506 468.2023 -211.0753 1 vs 2 27.32667   6e-04
```

```
anova(m2,m5)
```

```
##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m2     1 18 430.8239 477.7170 -197.4119
## m5     2 12 426.1938 457.4559 -201.0969 1 vs 2 7.369929   0.288
```

```
anova(m2,m6)
```

```
##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m2     1 18 430.8239 477.7170 -197.4119
## m6     2 10 442.1506 468.2023 -211.0753 1 vs 2 27.32667   6e-04
```

```
anova(m2,m7)
```

```
##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m2     1 18 430.8239 477.7170 -197.4119
## m7     2 13 427.3399 461.2071 -200.6700 1 vs 2 6.516009  0.2592
```

For m2 vs. m4 and m2 vs. m6, the results are significant so we compare the logLik and conclude that m2 is better. For m2 vs. m3, m2 vs. m5, m2 vs.m7, the results are non-significant and therefore we choose the reduced model, m3, m5 and m7.

Since compound symmetry (m3) is nested within the Toeplitz (m5) model, we conduct LRT and the result is not significant. Therefore, we choose the reduced model CS (m3).

```
anova(m3,m5)
```

```
##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m3     1 10 424.9320 450.9837 -202.4660
## m5     2 12 426.1938 457.4559 -201.0969 1 vs 2 2.738136  0.2543
```

Now we have m7 and m3 left. Since CS is nested within HCS and the LRT results are non-significant, we conclude that CS is the best model.

```
anova(m3,m7)
```

```
##    Model df      AIC      BIC   logLik   Test  L.Ratio p-value
## m3     1 10 424.9320 450.9837 -202.466
## m7     2 13 427.3399 461.2071 -200.670 1 vs 2 3.592055   0.309
```

iii. When choosing the best covariance best on statistical, show your work for selecting the best covariance for NOT nested models.

By the table we made, for AIC, m3 424.9320, m4 442.1506, m5 426.1938, m6 442.1506, m7 427.3399. So CS has the lowest AIC. Therefore, it's the best model.

```
print(fitstats)
```

```
##    covstr df      AIC    loglik
## m1    IND  9 481.8314 -231.9157
## m2     UN 18 430.8239 -197.4119
## m3     CS 10 424.9320 -202.4660
## m4  AR(1) 10 442.1506 -211.0753
## m5   Toep 12 426.1938 -201.0969
## m6    Exp 10 442.1506 -211.0753
## m7    CSH 13 427.3399 -200.6699
```

iv. Present your conclusion for the best covariance model.

Compound Symmetry is the best covariance model based on comparing the AIC and logLik using LRT.

Question 4 (10 points): note, this question is related to 3(b)(ii). Given the choice of model for the covariance from Question 3, treat age (or time) as a categorical variable and fit a model that includes the effects of age, gender, and their interactions.

The question asks for a pattern of change over time is different for boys and girls. Since we are looking at the pattern of change over time, this implies the interaction and not a gender effect.

(a) What type of a model is it for the mean response?

We use Compound Symmetry.

```
m3 <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID))
summary(m3)
```

```
## Generalized least squares fit by REML
##   Model: growth ~ cyear * gender
##   Data: dt.long
##       AIC      BIC   logLik
##   424.932 450.9837 -202.466
##
## Correlation Structure: Compound symmetry
##  Formula: ~time | ID
##  Parameter estimate(s):
##       Rho
## 0.6858426
##
## Coefficients:
##                     Value Std.Error  t-value p-value
## (Intercept)     21.181818 0.6688878 31.66722  0.0000
## cyear10          1.045455 0.5302026  1.97180  0.0514
## cyear12          1.909091 0.5302026  3.60068  0.0005
## cyear14          2.909091 0.5302026  5.48675  0.0000
## genderM          1.974432 0.8689107  2.27231  0.0252
## cyear10:genderM -0.389205 0.6887534 -0.56509  0.5733
## cyear12:genderM  0.653409 0.6887534  0.94868  0.3451
## cyear14:genderM  1.403409 0.6887534  2.03761  0.0442
##
##  Correlation:
##                 (Intr) cyer10 cyer12 cyer14 gendrM cy10:M cy12:M
```

```
## cyear10         -0.396
## cyear12         -0.396  0.500
## cyear14         -0.396  0.500  0.500
## genderM         -0.770  0.305  0.305  0.305
## cyear10:genderM  0.305 -0.770 -0.385 -0.385 -0.396
## cyear12:genderM  0.305 -0.385 -0.770 -0.385 -0.396  0.500
## cyear14:genderM  0.305 -0.385 -0.385 -0.770 -0.396  0.500  0.500
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.11040071 -0.66206139 -0.08195731  0.61307909  2.38060371
##
## Residual standard error: 2.21845
## Degrees of freedom: 108 total; 100 residual
```

(b) Determine whether the pattern of changes over time is different between boys and girls:

i. Use appropriate statistical test;

```
# Wald test
anova(m3,type = "marginal")
```

```
## Denom. DF: 100
##              numDF   F-value p-value
## (Intercept)      1 1002.8129  <.0001
## cyear            3   10.9205  <.0001
## gender           1    5.1634  0.0252
## cyear:gender     3    2.6037  0.0561
```

```
# LRT
m3_full <- gls(growth ~ cyear * gender,
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID),method='ML')
m3_reduced <- gls(growth ~ cyear + gender,
          data=dt.long,
          corr=corCompSymm(form= ~ time | ID),method='ML')
anova(m3_full,m3_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m3_full        1 10 426.6782 453.4995 -203.3391
## m3_reduced     2  7 428.7033 447.4782 -207.3517 1 vs 2 8.025105  0.0455
```

ii. Describe in words what you have found. Imagine that you are writing a conclusion in an abstract for a manuscript – use a couple of sentences at most.

The Compound Symmetry model is the best in our case. We conducted Likelihood Ratio Test as well as Wald Test to test the the pattern of change over time which is the interation effect between age and gender. LRT ($p = 0.0455$) shows that the presence of a significant interaction indicates that the effect of age variable on the response variable is different between different gender. Wald Test ($p = 0.0561$) shows an almost significant interation, however, the p-value is still greater than 0.05. It can be explained by that LRT is more conservative and that is known in small sample size.