# BIS628 HW1

*Joanna Chen*

*1/21/2020*

### Data Description

Dataset: cholesterol-data.txt
Each row of the data contains the following seven variables (wide format): Group ID Yi0 Yi6 Yi12 Yi20 Yi24
ID: patient ID number
Group: 1=High-dose treatment, 2=Placebo
Yij: cholesterol levels at baseline (j=0), 6 months (j=6), 12 months (j=12), 20 months (j=20), and 24 months (j=24)

### Data Cleaning

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

setwd("~/Downloads/BIS628HW1")
dt = read.csv("cholesterol-data.txt", header = F, sep = "")
colnames(dt) = c("group","ID","Y_i0","Y_i6","Y_i12","Y_i20","Y_i24")

dt.long = reshape2::melt(dt, id.vars = c('ID','group')) # reshape the data
setDT(dt.long)
setnames(dt.long,"value","serum") # rename the column

dt.long[,cgroup:=ifelse(group==2,0,1)] # create cgroup variable which treats group as a dummy variable
dt.long$cgroup = factor(dt.long$cgroup, levels=c(0,1), labels=c("Placebo","High-dose treatment"))

dt.long$month = substring(dt.long$variable, 4) # extract month
dt.long[,variable:=NULL] # delete the variable column since we already extract month from it
dt.long$cmonth = factor(dt.long$month, levels = c(0,6,12,20,24)) # create cmonth variable which treats 

dt.long$serum = as.numeric(dt.long$serum)

dt.long = dt.long[is.na(dt.long$serum)==F, ] # remove the missing value for the response variable
```

### Question 1

### (a). On a single graph, plot boxplots of mean cholesterol level across the follow up (in months, starting from the baseline) by group.

```
p1 = ggplot(data=dt.long, aes(y=serum, x=cmonth, fill=cgroup)) + geom_boxplot() +
  labs(title="Mean cholesterol level across the follow up over time by treatment group", x="Month Numbe
p1
```

### (a).a. Describe the variability in the outcome over time: report estimated variances by treatment group.

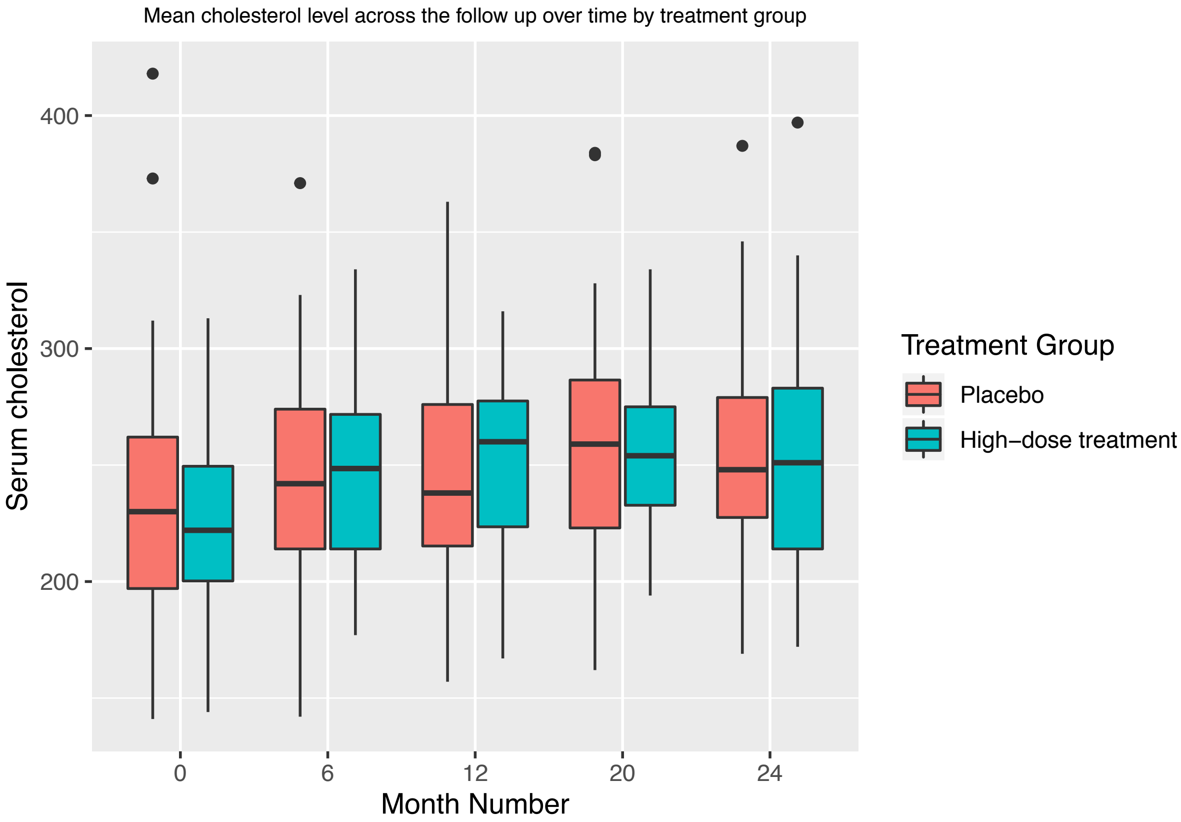

Figure 1: Mean cholesterol level across the follow up over time by treatment group

In the placebo group, the variance decreases from month 0 to month 12, then increases at month 20, then decreases again at month 24.In the high-dose treatment group, the variance decreases from month 0 to month 20, then significantly increases at month 24.

```
variance = dt.long %>%
  group_by(cgroup, cmonth) %>%
  summarize(var_serum = var(serum, na.rm = TRUE))
knitr::kable(variance)
```

| cgroup | cmonth | var_serum |
|---|---|---|
| Placebo | 0 | 3121.970 |
| Placebo | 6 | 2424.545 |
| Placebo | 12 | 2126.186 |
| Placebo | 20 | 2615.482 |
| Placebo | 24 | 2439.191 |
| High-dose treatment | 0 | 1573.262 |
| High-dose treatment | 6 | 1556.483 |
| High-dose treatment | 12 | 1469.129 |
| High-dose treatment | 20 | 1189.515 |
| High-dose treatment | 24 | 2496.200 |

### (b). On a Single graph, plot the observed mean serum cholesterol versus time (in months, starting from the baseline) by group

```
mean = dt.long %>%
  group_by(cgroup, cmonth) %>%
  summarize(mean_serum = mean(serum, na.rm = TRUE))
# mean

p2 = ggplot(mean, aes(x = cmonth, y = mean_serum, color = cgroup, group = cgroup )) +
  geom_point() +
  geom_line() +
  labs(title="Observed mean serum cholesterol versus time",
       x="Month", y="Serum cholesterol", color = "cgroup") +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
p2
```

### (b)a. Describe the general characteristics of the time trends for the two groups: what is going on with the group means over time?

Both groups' serum cholesterol mean are increasing from month 0 to month 20, and both are decreasing at month 24. At the baseline month 0, the serum cholesterol mean of high-dose treatement group is less than Placebo group. At the month 6 and 12, the serum cholesterol mean of high-dose treatement group is higher than Placebo group, At month 20, both mean are very close to each other, while at the month 0 and 24, the mean of treatment group is less than Placebo group again.

## 2. Assuming an unstructured covariance matrix, conduct an analysis of response profiles. Determine whether the patterns of change over time differ in the two treatment groups:

### a. With baseline (month 0) and the placebo group (group 2) as the reference groups, using the dummy coding approach, write out the regression model for the mean serum cholesterol: $E(Y_{ij})$

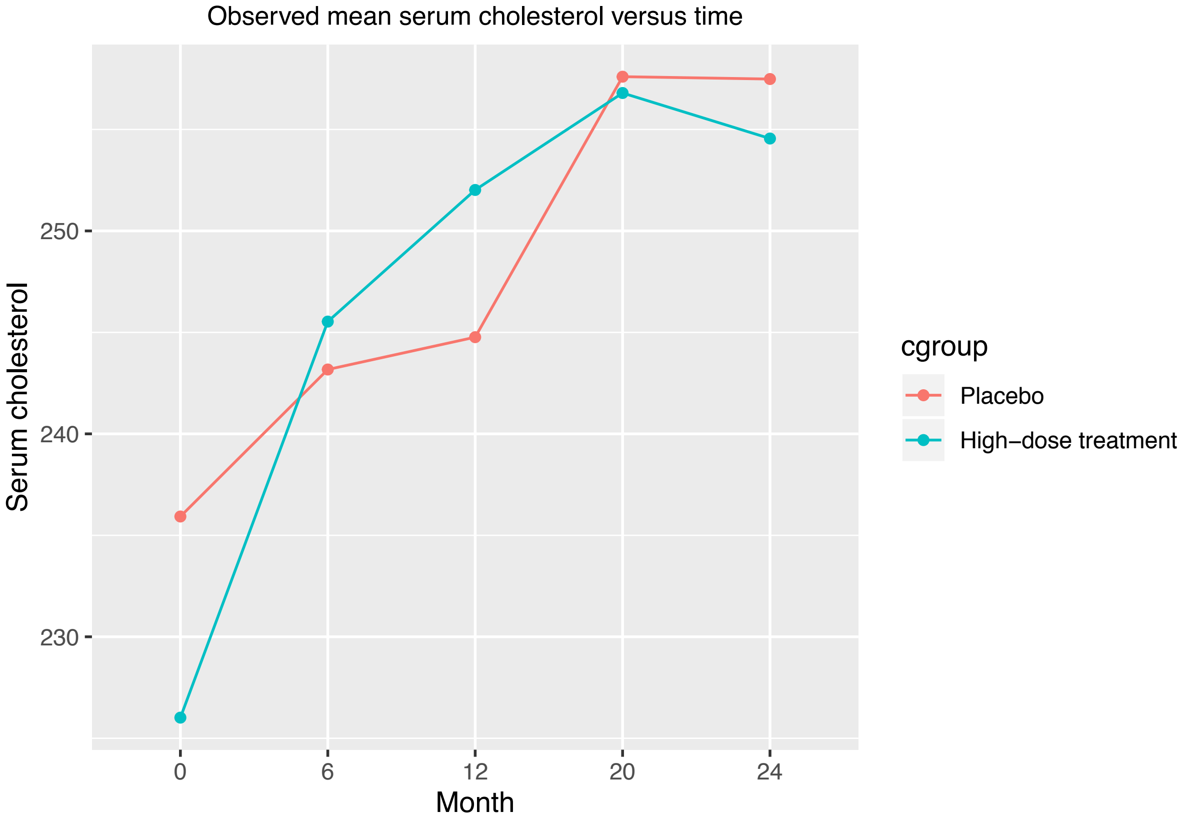

Figure 2: Observed mean serum cholesterol versus time

$$E(Y_{ij}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9$$

where $\beta_0$ = intercept when baseline time is baseline 0 and group is placebo
$\beta_1$ = group effect from high dose group
$\beta_2$ = time effect from month 6
$\beta_3$ = time effect from month 12
$\beta_4$ = time effect from month 20
$\beta_5$ = time effect from month 24
$\beta_6$ = time effect from month 6 and high dose group
$\beta_7$ = time effect from month 12 and high dose group
$\beta_8$ = time effect from month 20 and high dose group
$\beta_9$ = time effect from month 24 and high dose group

$$X_1 = \begin{cases} 1 & \text{group = high dose} \\ 0 & \text{otherwise (placebo)} \end{cases}$$
$$X_2 = \begin{cases} 1 & \text{time = month 6} \\ 0 & \text{otherwise} \end{cases}$$
$$X_3 = \begin{cases} 1 & \text{time = month 12} \\ 0 & \text{otherwise} \end{cases}$$
$$X_4 = \begin{cases} 1 & \text{time = month 20} \\ 0 & \text{otherwise} \end{cases}$$
$$X_5 = \begin{cases} 1 & \text{time = month 24} \\ 0 & \text{otherwise} \end{cases}$$

$X_6$= $X_1 X_2$ = interation between time = month 6 and group = high dose
$X_7$= $X_1 X_3$ = interation between time = month 12 and group = high dose
$X_8$= $X_1 X_4$ = interation between time = month 20 and group = high dose
$X_9$= $X_1 X_5$ = interation between time = month 24 and group = high dose

**b. Write out the mean response model for subjects in the Placebo group: $\mu ij$(Placebo)**

Let $X_1$ = group = 0. Then

$$\begin{aligned} \mu_{ij}(Placebo) = E(Y_{ij,placebo}) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1 X_2 + \beta_7 X_1 X_3 + \beta_8 X_1 X_4 + \beta_9 X_1 X_5 \\ &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \end{aligned}$$

**c. Write out the mean response model for subject in the High-dose group: $\mu ij$(High-Dose)**

Let $X_1$ = group = 1. Then

$$\begin{aligned} \mu_{ij}(High-Dose) = E(Y_{ij,high-dose}) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1 X_2 + \beta_7 X_1 X_3 + \beta_8 X_1 X_4 + \beta_9 X_1 \\ &= \beta_0 + \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_2 + \beta_7 X_3 + \beta_8 X_4 + \beta_9 X_5 \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_6) X_2 + (\beta_3 + \beta_7) X_3 + (\beta_4 + \beta_8) X_4 + (\beta_5 + \beta_9) X_5 \end{aligned}$$

**d. Implement the analysis of response profiles in SAS or R:**

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
## The following object is masked from 'package:dplyr':
## 
##     collapse
```

```
# https://stackoverflow.com/questions/12925063/numbering-rows-within-groups-in-a-data-frame
dt.long = dt.long %>% group_by(ID) %>% mutate(time = row_number())

model <- gls(serum ~ cgroup*cmonth,
             data = dt.long,
             corr = corSymm(form = ~ time | ID),
             weights = varIdent(form = ~ 1 | month),
             na.action = na.omit
            )
summary(model)
```

```
## Generalized least squares fit by REML
##   Model: serum ~ cgroup * cmonth 
##   Data: dt.long 
##        AIC      BIC    logLik
##   4314.588 4416.587 -2132.294
## 
## Correlation Structure: General
##  Formula: ~time | ID 
##  Parameter estimate(s):
##  Correlation: 
##   1     2     3     4    
## 2 0.770                  
## 3 0.732 0.773            
## 4 0.738 0.800 0.726      
## 5 0.586 0.665 0.678 0.625
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | month 
##  Parameter estimates:
##         0         6        12        20        24 
## 1.0000000 0.9320568 0.8791668 0.8974016 1.0300809 
## 
## Coefficients:
##                                         Value Std.Error  t-value p-value
## (Intercept)                         235.92683  7.305948 32.29243  0.0000
## cgroupHigh-dose treatment            -9.67829  9.412956 -1.02819  0.3044
## cmonth6                               7.24390  4.805425  1.50744  0.1324
## cmonth12                              8.84620  5.207262  1.69882  0.0901
## cmonth20                             23.10333  5.292171  4.36557  0.0000
## cmonth24                             21.12230  7.398137  2.85508  0.0045
## cgroupHigh-dose treatment:cmonth6    12.21751  6.193407  1.97266  0.0492
## cgroupHigh-dose treatment:cmonth12  16.28893  6.738391  2.41733  0.0160
## cgroupHigh-dose treatment:cmonth20   4.75670  6.973253  0.68213  0.4955
## cgroupHigh-dose treatment:cmonth24   6.53598  9.763271  0.66945  0.5036
## 
##  Correlation: 
##                                    (Intr) cgrH-t cmnth6 cmnt12 cmnt20
## cgroupHigh-dose treatment          -0.776                            
## cmonth6                            -0.429  0.333                     
## cmonth12                           -0.500  0.388  0.581              
## cmonth20                             -0.466  0.362  0.606  0.526
## cmonth24                             -0.392  0.304  0.476  0.522  0.438
## cgroupHigh-dose treatment:cmonth6     0.333 -0.429 -0.776 -0.451 -0.470
## cgroupHigh-dose treatment:cmonth12  0.387 -0.497 -0.449 -0.773 -0.407
## cgroupHigh-dose treatment:cmonth20  0.354 -0.456 -0.460 -0.400 -0.759
## cgroupHigh-dose treatment:cmonth24  0.297 -0.378 -0.361 -0.396 -0.332
##                                     cmnt24 cH-t:6 cH-t:1 cH-t:20
## cgroupHigh-dose treatment
## cmonth6
## cmonth12
## cmonth20
## cmonth24
## cgroupHigh-dose treatment:cmonth6  -0.369
## cgroupHigh-dose treatment:cmonth12 -0.404  0.578
## cgroupHigh-dose treatment:cmonth20 -0.332  0.592  0.513
## cgroupHigh-dose treatment:cmonth24 -0.758  0.463  0.503  0.419
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.32029916 -0.68866948 -0.02685013  0.60855779  3.89204113
##
## Residual standard error: 46.7809
## Degrees of freedom: 447 total; 437 residual
```

### i. Output an estimated variance/covariance matrix and the correlation matrix:

The corvariance matrix is the first matrix of the output. The variance matrix is the last matrix of the output. The things in the middle is just weights to compute variance.

```
# The code in this chunk is from Lab 2
#Function to Extract and Print Variance/covariance matrix from 'gls' function
corandcov <- function(glsob,cov=T,...){
corm <- corMatrix(glsob$modelStruct$corStruct)[[5]]
print(corm)
varstruct <- print(glsob$modelStruct$varStruct)
varests <- coef(varstruct, uncons=F, allCoef=T)
covm <- corm*glsob$sigma^2*t(t(varests))%*%t(varests)
return(covm)}

#Print Var/Covariance matrix from model
corandcov(model)
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.7703962 0.7317579 0.7379188 0.5858649
## [2,] 0.7703962 1.0000000 0.7734828 0.7996121 0.6651360
## [3,] 0.7317579 0.7734828 1.0000000 0.7264887 0.6777836
## [4,] 0.7379188 0.7996121 0.7264887 1.0000000 0.6248963
## [5,] 0.5858649 0.6651360 0.6777836 0.6248963 1.0000000
## Variance function structure of class varIdent representing
##         0         6        12        20        24
## 1.0000000 0.9320568 0.8791668 0.8974016 1.0300809
```

```
##          0        6       12       20       24
## 0  2188.452 1571.425 1407.913 1449.214 1320.705
## 6  1571.425 1901.174 1387.080 1463.678 1397.530
## 12 1407.913 1387.080 1691.530 1254.365 1343.293
## 20 1449.214 1463.678 1254.365 1762.425 1264.163
## 24 1320.705 1397.530 1343.293 1264.163 2322.094
```

### 2i1. Comment on the variances and covariances, as well as the correlations for pairs of responses ($Y_{ij}$,$Y_{ij'}$)

Variance is on the diagonal - $(2188, 1901, 1692, 1762, 2322)$. It decreases from month 0 to month 12, then increases from month 12 to month 24. Month 24 has the largest variance. We can see the off-diagnoal terms of the triangular matrix in our variancn-covariance output that the covariance between month 0 and other months overall decreases from month 0 to month 24. Specifically, it's monotonically decreasing until month 12, increase a little bit at month 20, and then decrease again at month 24. The covariance of month 6 decreases at month 12, increase at month 20 and decrease again at month 24. The covariance of month 12 increases from 1250 at month 20 to 1347 at month 24. The covariance of month 20 and 24 is 1264.
The mean response profile analysis estimated correlation that are 6,6,8,4 month apart of 0.77,0.77,0.73,0.62. For 12 months apart the correlation decreases slightly to 0.73 and 0.67. Finally, correlation that 24 months apart is of 0.58. Thus, the correlation tends to decrease with increasing time separation between the measurement times. Moreover, the larger varibility at month 24 translates into lower correlation with other months.

### 2i2. Conduct an appropriate hypothesis test for the differences in mean serum cholesterol levels over time between the two groups.

#### 2i2a. Use an omnibus test (Wald test) (can use likelihood ratio test (LRT))

Here I just write out the Wald test hypothesis. I am going to use LRT to do it.
For Wald test:

$$H_0 : L\beta = 0, H_a : L\beta \neq 0$$

where

$$\beta = (\beta_0\beta_1\beta_2\beta_3\beta_4\beta_5\beta_6\beta_7\beta_8\beta_9)^T$$

and

$$L = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

For likelikehood ratio test,

$$H_0 : \beta_6 + \beta_7 + \beta_8 + \beta_9 = 0, H_a : \beta_6 + \beta_7 + \beta_8 + \beta_9 \neq 0$$

```
m_full <- gls(serum ~ cgroup*cmonth, # note that, cgroup*cmonth here = cgroup + cmonth + cgroup*cmonth
          data = dt.long,
          corr = corSymm(form= ~ time | ID),
          weights = varIdent(form= ~ 1 | cmonth),
          na.action = na.omit,
          method='ML') #the default is REML, so we need to specify ML here
m_reduced <- gls(serum ~ cgroup + cmonth,
          data = dt.long,
          corr = corSymm(form= ~ time | ID),
          weights = varIdent(form= ~ 1 | cmonth),
          na.action = na.omit,
          method='ML')

anova(m_full, m_reduced) #compare the two models (LR test)
```

```
##           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m_full        1 25 4363.233 4465.797 -2156.616
## m_reduced     2 21 4362.957 4449.111 -2160.479 1 vs 2 7.724584  0.1022
```

The p-value is $0.1022 > 0.05$, therefore we fail to reject the null hypothesis. We conclude that there are no differences in changes of mean serum cholesterol levels over time between the two groups.

*Note to myself*: An interesting question to think about - We know that the mean difference between two groups is $\mu ij(high - dose) - \mu ij(placebo) = \beta_1 + \beta_6 + \beta_7 + \beta_8 + \beta_9$. Why the null hypothesis is not

$$H_0 : \beta_1 + \beta_6 + \beta_7 + \beta_8 + \beta_9 = 0.$$

This is becuase $\beta_1$ is a constant parameter, while we would like to test the parameter that varies.

### 2i2b. Which regression parameters (estimated coefficients) test for the group differences at each time point?

$\beta_6$ test for group difference at month 6
$\beta_7$ test for group difference at month 12
$\beta_8$ test for group difference at month 20
$\beta_9$ test for group difference at month 24

### 2i3. Using the estimated regression coefficients from the analysis of response profiles, construct time-specific estimated means for the following:

Plug in the $\beta$'s, we get

$$\begin{aligned} E(Y_{ij}) = & -9.68 \times group + 7.24 \times month6 + 8.85 \times month12 + 23.10 \times month20 \\ & + 21.12 \times month24 + 12.22 \times group \times month6 + 16.29 \times group \times month12 \\ & + 4.76 \times group \times month20 + 6.54 \times group \times month24 + 235.93 \end{aligned}$$

### 2ia. Estimated Mean in the Placebo Group at month 12

$$\begin{aligned} \mu ij(placebo)\text{at month 12} &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \\ &= 7.24 \times month6 + 8.85 \times month12 + 23.10 \times month20 + 21.12 \times month24 + 235.93 \\ &= 7.24 \times 0 + 8.85 \times 1 + 23.10 \times 0 + 21.12 \times 0 + 235.93 \\ &= 8.85 + 235.93 \\ &= 244.78 \end{aligned}$$

### 2ib. Mean in the Treatment (High-dose) Group at month 20

$$\begin{aligned}\mu ij(highdose)\text{at month 20} &= (\beta_0+\beta_1)+(\beta_2+\beta_6)X_2+(\beta_3+\beta_7)X_3+(\beta_4+\beta_8)X_4+(\beta_5+\beta_9)X_5 \\ &= (7.24+12.22)\times month6+(8.85+16.29)\times month12+(23.10+4.76)\times month20 \\ &+ (21.12+6.54)\times month24+(235.93-9.68) \\ &= 19.46\times month6+25.14\times month12+27.86\times month20+27.66\times month24+226.25 \\ &= 19.46\times 0+25.14\times 0+27.86\times 1+27.66\times 0+226.25 \\ &= 254.11\end{aligned}$$

**2e. Plot the observed and the estimated from this model mean serum cholesterol levels for each group (connect the means to produce mean profiles, i.e. connect the means to obtain jagged lines):**

```
#Extract unique combinations of cgroup and cmonth and predicted values of serum ( means)
p.model <- unique(data.frame(cgroup=dt.long$cgroup,cmonth=dt.long$cmonth,serum=model$fitted))
p.model$group <- "Predicted"

#Calculate group means of serum at each occasion
dt.long_rm_missing_subset <- aggregate(dt.long$serum, list(dt.long$cgroup, dt.long$cmonth), mean)
#Rename the variables in dt.long_rm_missing_subset
names(dt.long_rm_missing_subset) <- c("cgroup","cmonth","serum")
#Sort it by treatment group
dt.long_rm_missing_subset <- dt.long_rm_missing_subset[order(dt.long_rm_missing_subset$cgroup),]
dt.long_rm_missing_subset$group <- "Observed"

#combine the dataframe with dt.long_rm_missing by row
p.model <- rbind(p.model,dt.long_rm_missing_subset)

p3 <- ggplot(p.model, aes(x=cmonth,y=serum,color=cgroup,linetype=group,group=interaction(cgroup,group)))
p3 + geom_line() + geom_point() +
  labs(title="Observed and estimated mean serum cholesterol levels by groups from the mean response mode
       x="Month Number", y="Serum cholesterol", color="Treatment Group", linetype="Predicted/Observed")
  scale_linetype_manual(values=c("dashed","solid"), labels=c("Observed","Predicted")) +
  theme(plot.title = element_text(hjust = 0.5,size = 9))
```

**2i. What are your observations?**

The predicted mean serum cholesterol levels for each group from the mean response model is close to the observed value and follows the pattern of the observed value, although there are some error. The prediction of high-dose treatement is under-estimated. Before month 20, both groups mean serum cholesterol levels increases, while after month 20, both decreases.

**Question 3: Assuming an unstructured covariance matrix, model the mean levels of serum cholesterol using a linear curve. Determine whether the curves differ between the two treatment groups:**

**a. With baseline (month 0) and the placebo group (group 2) as the reference groups, using the dummy coding approach, write out the regression model for the mean serum cholesterol:** $E(Y_{ij})$

$$E(Y_{ij})=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3$$

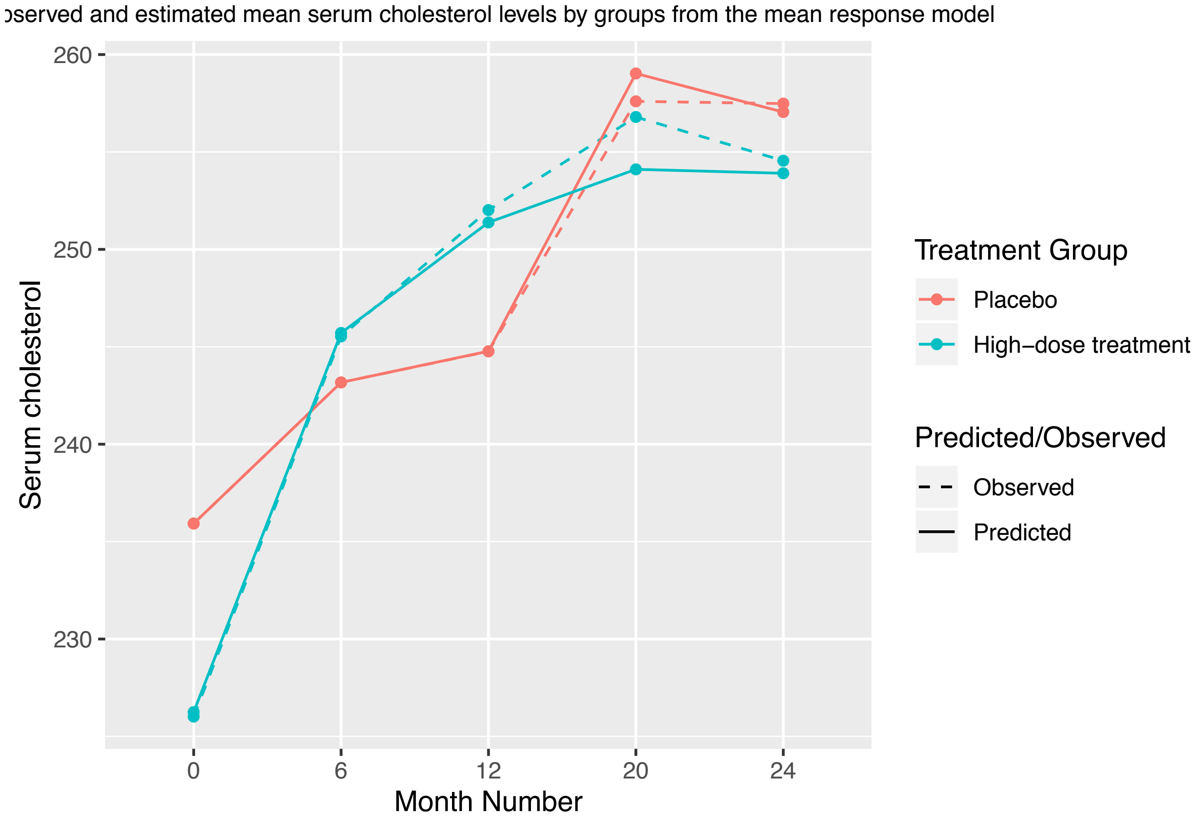

Figure 3: Observed and estimated mean serum cholesterol levels by groups from the mean response model

where $\beta_0$ = intercept when baseline time is at baseline month 0 and group is placebo
$\beta_1$ = time effect
$\beta_2$ = group effect
$\beta_3$ = interation of time and group effect

$$X_1 = time_{ij} = \text{measurement time for the jth measurement on ith individual.}$$
$$X_2 = group_i = \begin{cases} 1 & \text{if the ith individual was assigned to the high-dose group} \\ 0 & \text{otherwise} \end{cases}$$
$$X_3 = time_{ij} \times group_i$$

### b. Write out the mean response model for subjects in the Placebo group: $\mu_{ij}(Placebo)$

For placebo group, $X_2 = 0$ and therefore $X_3 = 0$. Hence,

$$\mu_{ij}(Placebo) = E(Y_{ij}, placebo) = \beta_0 + \beta_1 X_1$$

### c. Write out the mean response model for subject in the High-dose group: $\mu_{ij}(High - Dose)$

For high-dose group, $X_2 = 1$ and therefore $X_3 = X_1$. Hence,

$$\mu_{ij}(High - Dose) = E(Y_{ij}, high - dose) = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$

### d. Implement the analysis in SAS or R:

```
dt.long$month = as.numeric(dt.long$month)
dt.long$time = as.numeric(dt.long$time)
dt.long$month = as.integer(dt.long$month)

model2 <- gls(serum ~ cgroup*month,
              # use cgroup rather than group. there are different when number of group >=3.
              data = dt.long,
              corr = corSymm(form = ~ time | ID),
              weights = varIdent(form = ~ 1 | cmonth),
              na.action = na.omit
)

summary(model2)
```

```
## Generalized least squares fit by REML
##   Model: serum ~ cgroup * month
##   Data: dt.long
##        AIC      BIC    logLik
##   4360.437 4438.215 -2161.218
##
## Correlation Structure: General
##  Formula: ~time | ID
##  Parameter estimate(s):
##  Correlation:
##   1     2     3     4
## 2 0.753
## 3 0.716 0.775
## 4 0.737 0.793 0.718
## 5 0.589 0.652 0.676 0.632
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | cmonth
##  Parameter estimates:
##         0         6        12        20        24
## 1.0000000 0.9249446 0.8734078 0.8869052 1.0253078
##
## Coefficients:
##                                    Value Std.Error  t-value p-value
## (Intercept)                    235.21912  6.768419 34.75245  0.0000
## cgroupHigh-dose treatment       -0.36345  8.729531 -0.04163  0.9668
## month                            1.02238  0.218884  4.67088  0.0000
## cgroupHigh-dose treatment:month  0.07559  0.289137  0.26144  0.7939
##
##  Correlation:
##                                 (Intr) cgrH-t month
## cgroupHigh-dose treatment       -0.775
## month                           -0.459  0.356
## cgroupHigh-dose treatment:month  0.348 -0.455 -0.757
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.27712220 -0.71518301 -0.04930373  0.63412698  3.87480752
##
## Residual standard error: 47.1716
## Degrees of freedom: 447 total; 443 residual
```

### ii. Output an estimated variance/covariance matrix and the correlation matrix:

The corvariance matrix is the first matrix of the output. The variance matrix is the last matrix of the output. The things in the middle is just weights to compute variance.

```
corandcov(model2)
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.7525869 0.7161332 0.7366201 0.5890061
## [2,] 0.7525869 1.0000000 0.7748208 0.7933064 0.6518027
## [3,] 0.7161332 0.7748208 1.0000000 0.7182353 0.6761054
## [4,] 0.7366201 0.7933064 0.7182353 1.0000000 0.6318200
## [5,] 0.5890061 0.6518027 0.6761054 0.6318200 1.0000000
## Variance function structure of class varIdent representing
##         0         6        12        20        24
## 1.0000000 0.9249446 0.8734078 0.8869052 1.0253078
```

```
##           0        6       12       20       24
## 0  2225.160 1548.937 1391.785 1453.724 1343.802
## 6  1548.937 1903.675 1392.821 1448.089 1375.458
## 12 1391.785 1392.821 1697.444 1238.004 1347.246
## 20 1453.724 1448.089 1238.004 1750.313 1278.457
## 24 1343.802 1375.458 1347.246 1278.457 2339.213
```

### 1. Comment on the variances and covariance, as well as the correlations for pairs of responses (Yij,Yij’)

Variance is on the diagonal - $(2225, 1903, 1697, 1750, 2339)$. It decreases from month 0 to month 12, then increases from month 12 to month 24. Month 24 has the largest variance. We can see the off-diagnoal terms of the triangular matrix in our variance-covariance output that the covariance between month 0 and other months overall decreases from month 0 to month 24. Specifically, it's monotonically decreasing until month 12, increase a little bit at month 20, and then decrease again at month 24. The covariance of month 6 decreases at month 12, increase at month 20 and decrease again at month 24. The covariance of month 12 increases from 1238 at month 20 to 1347 at month 24. The covariance of month 20 and 24 is 1278.
The mean response profile analysis estimated correlation that are 6,6,8,4 month apart of 0.75,0.77,0.72,0.63. For 12 months apart the correlation decreases slightly to 0.72 and 0.68. Finally, correlation that 24 months apart is of 0.59. Thus, the correlation tends to decrease with increasing time separation between the measurement times. Moreover, the larger varibility at month 24 translates into lower correlation with other months.

### 2. Conduct an appropriate hypothesis test for the differences in mean serum cholesterol levels over time between the two groups.

#### a. Use the Wald test (can use LRT)

Here, we use likelikehood ratio test. The null hypothesis is that the interation's parameter equals zero.

$$H_0 : \beta_3 = 0, H_a : \beta_3 \neq 0$$

```
m2_full <- gls(serum ~ cgroup*month,
               data = dt.long,
               corr = corSymm(form= ~ time | ID),
               weights = varIdent(form= ~ 1 | cmonth),
               na.action = na.omit,
               method='ML')

m2_reduced <- gls(serum ~ cgroup + month,
                  data = dt.long,
                  corr = corSymm(form= ~ time | ID),
                  weights = varIdent(form= ~ 1 | cmonth),
                  na.action = na.omit,
                  method='ML')

#Compare the two models (LR test)
anova(m2_full, m2_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test   L.Ratio p-value
## m2_full        1 19 4368.149 4446.098 -2165.075
## m2_reduced     2 18 4366.210 4440.056 -2165.105 1 vs 2 0.0604136  0.8058
```

The p-value is $0.8058 > 0.05$, therefore we fail to reject the null hypothesis. We conclude that there are no differences in patterns of change of mean serum cholesterol levels over time between the two groups.

#### b. Which regression parameters (estimated coefficient(s)) test for the group differences over time?

$\beta_3$ test for group difference over time.

### 3. Using the estimated regression coefficients from the analysis of response profiles, construct time-specific estimated means for the following:

### a. Estimated Mean in the Placebo Group at month 12

$$\mu_{ij}(Placebo) \text{ at month } 12 = \beta_0 + \beta_1 X_1 = \beta_0 + 12\beta_1 = 235.22 + 12 \times 1.02 = 247.46$$

### b. Mean in the Treatment (High-dose) Group at month 20

$$\mu_{ij}(High{-}Dose) \text{ at month } 12 = (\beta_0+\beta_2)+(\beta_1+\beta_3)X_1 = (\beta_0+\beta_2)+20(\beta_1+\beta_3) = 235.22-0.36+20\times(1.02+0.08) = 256.86$$

### e. Plot the observed and predicted from this model mean serum cholesterol levels for each group over time (these should be lines):

```r
#Extract unique combinations of group and month and predicted values of serum (means)
p.model2 <- unique(data.frame(cgroup=dt.long$cgroup,cmonth=dt.long$cmonth,serum=model2$fitted))

p.model2$group <- "Predicted"

p.model2 <- rbind(p.model2,dt.long_rm_missing_subset)

p4 <- ggplot(p.model2, aes(x=as.numeric(as.character(cmonth)),y=serum,color=cgroup,linetype=group,group=
       x="Month Number", y="Serum cholesterol", color="Treatment Group", linetype="Predicted/Observed")
  scale_linetype_manual(values=c("dashed","solid"), labels=c("Observed","Predicted")) +
  theme(plot.title = element_text(hjust = 0.5,size = 8))
p4
```

### iii. What are your observations?

The predicted mean serum cholesterol levels for each group from the mean response model is quite far from the observed value and the residual is larger than the previous mean response model. Comparing with two groups, the placebo model prediction looks better than the high-dose treatement's. At month 0 and 24, the high-dose treatement was over estimated, while underestimated at 6 and 12. The placebo group at month 12 was overestimated.

### Question 4. Assuming an unstructured covariance matrix, model the mean levels of serum cholesterol using a piecewise linear spline with a knot at 12 months. Determine whether the splines differ between the two treatment groups:

### f. With baseline (month 0) and the placebo group (group 2) as the reference groups, using the dummy coding approach, write out the regression model for the mean serum cholesterol: E(Yij)

$$E(Y_{ij}) = \beta_1 + \beta_2 X_1 + \beta_3 (X_1 - t^*)_+ + \beta_4 X_2 + \beta_5 X_1 X_2 + \beta_6 (X_1 - t^*)_+ X_2 \text{ where } t^* = 12$$

$$X_1 = time_{ij} = \text{measurement time for the jth measurement on ith individual.}$$
$$X_2 = group_i = \begin{cases} 1 & \text{if the ith individual was assigned to the high-dose group} \\ 0 & \text{otherwise} \end{cases}$$

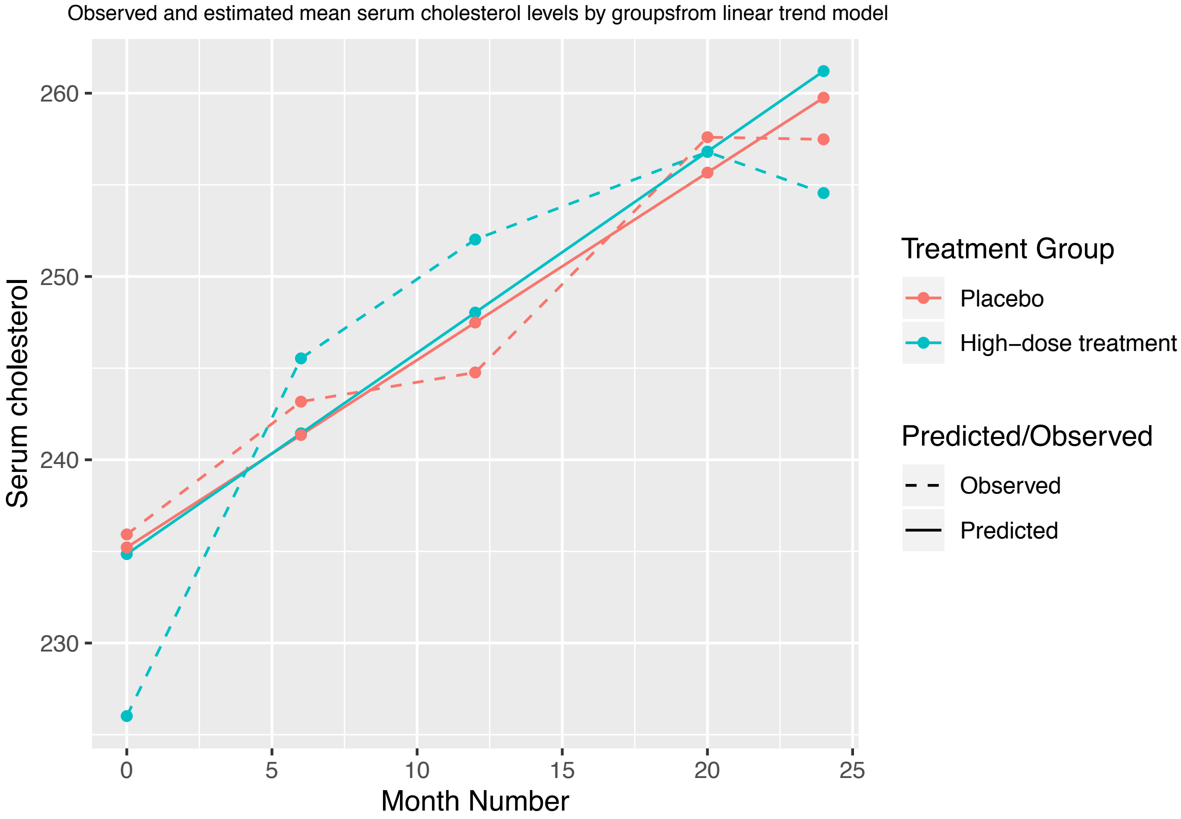

Figure 4: Observed and estimated mean serum cholesterol levels by groups from linear trend model

**g. Write out the mean response model for subjects in the Placebo group:** $\mu_{ij}(Placebo)$

Let $X_2 = 0$. Then

$$\mu_{ij}(Placebo) = \beta_1 + \beta_2 X_1 + \beta_3 (X_1 - t^*)_+$$

At month 12 knot:

$$X_1 \leq 12month : \mu_{ij}(Placebo) = \beta_1 + \beta_2 X_1$$

$$X_1 > 12month : \mu_{ij}(Placebo) = \beta_1 + \beta_2 X_1 + \beta_3 (X_1 - 12)_+ = (\beta_1 - 12\beta_3) + (\beta_2 + \beta_3) X_1$$

**h. Write out the mean response model for subject in the High-dose group:** $\mu_{ij}(High - Dose)$

Let $X_2 = 1$. Then

$$\mu_{ij}(High{-}Dose) = \beta_1 + \beta_2 X_1 + \beta_3 (X_1 - t^*)_+ + \beta_4 + \beta_5 X_1 + \beta_6 (X_1 - t^*)_+ = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) X_1 + (\beta_3 + \beta_6)(X_1 - t^*)_+$$

At month 12 knot,

$$X_1 \leq 12month : \mu_{ij}(High - dose) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) X_1$$

$$X_1 > 12month : \mu_{ij}(High{-}dose) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) X_1 + (\beta_3 + \beta_6)(X_1 - 12)_+ = (\beta_1 + \beta_4) - 12(\beta_3 + \beta_6) + (\beta_2 + \beta_3 + \beta_5 + \beta_6) X_1$$

**i. Implement the analysis in SAS or R:**

```r
#Create a spline variable
dt.long$month_12 <- (dt.long$month-12)*I(dt.long$month>=12)

### Fit a piecewise linear model
model4 <- gls(serum ~ group*month+group*month_12,
              data=dt.long,
              corr=corSymm(form= ~ time | ID),
              weights=varIdent(form= ~ 1 | cmonth),
              na.action = na.omit)
summary(model4)
```

```
## Generalized least squares fit by REML
##   Model: serum ~ group * month + group * month_12
##   Data: dt.long
##        AIC      BIC    logLik
##   4351.582 4437.452 -2154.791
##
## Correlation Structure: General
##  Formula: ~time | ID
##  Parameter estimate(s):
##  Correlation:
##   1     2     3     4
## 2 0.766
## 3 0.732 0.769
## 4 0.736 0.800 0.725
## 5 0.586 0.666 0.677 0.625
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | cmonth
##  Parameter estimates:
##         0         6        12        20        24
## 1.0000000 0.9332588 0.8785994 0.8954334 1.0270880
##
## Coefficients:
##                    Value Std.Error   t-value p-value
## (Intercept)    221.19790 13.635358 16.222376  0.0000
## group            7.55959  9.204201  0.821320  0.4119
## month            3.28421  0.805946  4.074976  0.0001
## month_12        -4.59662  1.445704 -3.179502  0.0016
## group:month     -1.22139  0.540572 -2.259447  0.0243
## group:month_12   2.48535  0.960869  2.586566  0.0100
##
##  Correlation:
##                (Intr) group  month  mnt_12 grp:mn
## group          -0.944
## month          -0.496  0.468
## month_12        0.316 -0.297 -0.843
## group:month     0.471 -0.497 -0.944  0.795
## group:month_12 -0.301  0.317  0.802 -0.945 -0.848
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.27454509 -0.68736078 -0.01618232  0.63165265  3.88127641
##
## Residual standard error: 46.81009
## Degrees of freedom: 447 total; 441 residual
```

### iv. Output an estimated variance/covariance matrix and the correlation matrix:

```r
corandcov(model4)
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.7661800 0.7320661 0.7362993 0.5858915
## [2,] 0.7661800 1.0000000 0.7688014 0.7999444 0.6656245
## [3,] 0.7320661 0.7688014 1.0000000 0.7249662 0.6774498
## [4,] 0.7362993 0.7999444 0.7249662 1.0000000 0.6252491
## [5,] 0.5858915 0.6656245 0.6774498 0.6252491 1.0000000
## Variance function structure of class varIdent representing
##         0         6        12        20        24
## 1.0000000 0.9332588 0.8785994 0.8954334 1.0270880
```

```
##           0        6       12       20       24
## 0  2191.185 1566.794 1409.354 1444.664 1318.572
## 6  1566.794 1908.461 1381.294 1464.786 1398.035
## 12 1409.354 1381.294 1691.456 1249.744 1339.537
## 20 1444.664 1464.786 1249.744 1756.894 1260.007
## 24 1318.572 1398.035 1339.537 1260.007 2311.502
```

### 1. Comment on the variances and covariances, as well as the correlations for pairs of responses (Yij,Yij')

Variance is on the diagonal - $(2191, 1908, 1691, 1756, 2311)$. It decreases from month 0 to month 12, then increases from month 12 to month 24. Month 24 has the largest variance. We can see the off-diagnoal terms of the triangular matrix in our variance-covariance output that the covariance between month 0 and other months overall decreases from month 0 to month 24. Specifically, it's monotonically decreasing until month

12, increase at month 20, and then decrease again at month 24. The covariance of month 6 decreases at month 12, increases a little at month 20 and decrease again at month 24. The covariance of month 12 increases from 1249 at month 20 to 1339 at month 24. The covariance of month 20 and 24 is 1260.
The mean response profile analysis estimated correlation that are 6,6,8,4 month apart of 0.77,0.77,0.72,0.63. For 12 months apart the correlation decreases slightly to 0.73 and 0.68. Finally, correlation that 24 months apart is of 0.59. Thus, the correlation tends to decrease with increasing time separation between the measurement times. Moreover, the larger varibility at month 24 translates into lower correlation with other months.

### 2. Conduct an appropriate hypothesis test for the differences in mean serum cholesterol levels over time between the two groups.

#### a. Use the Wald test (a contrast statement)

For likelikehood ratio test,

$$H_0 : \beta_5 = \beta_6 = 0, H_a : \text{at least one of them is not zero}$$

```
m4_full <- gls(serum ~ group*month+group*month_12,
           data=dt.long,
           corr=corSymm(form= ~ time | ID),
           weights=varIdent(form= ~ 1 | cmonth),
           method='ML')
m4_reduced <- gls(serum ~ group+month+month_12,
           data=dt.long,
           corr=corSymm(form= ~ time | ID),
           weights=varIdent(form= ~ 1 | cmonth),
           method='ML')
anova(m4_full, m4_reduced)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m4_full        1 21 4361.343 4447.496 -2159.671
## m4_reduced     2 19 4363.814 4441.763 -2162.907 1 vs 2 6.471502  0.0393
```

The p-value is $0.0393 < 0.05$, therefore we reject the null hypothesis. We conclude that there are statistically significant differences in patterns of changes of mean serum cholesterol levels over time between the two groups.

#### b. Which regression parameters (estimated coefficient(s)) test for the group differences over time?

$\beta_5, \beta_6$

### 3. Using the estimated regression coefficients from the analysis of response profiles, construct time-specific estimated means for the following:

#### a. Estimated Mean in the Placebo Group at month 12

Since $X_1 = 12 \leq 12$ month, we have

$$\mu_{ij}(Placebo, \text{month } 12) = \beta_1 + \beta_2 X_1 = \beta_1 + 12\beta_2 = 221.20 + 3.28 = 224.48$$

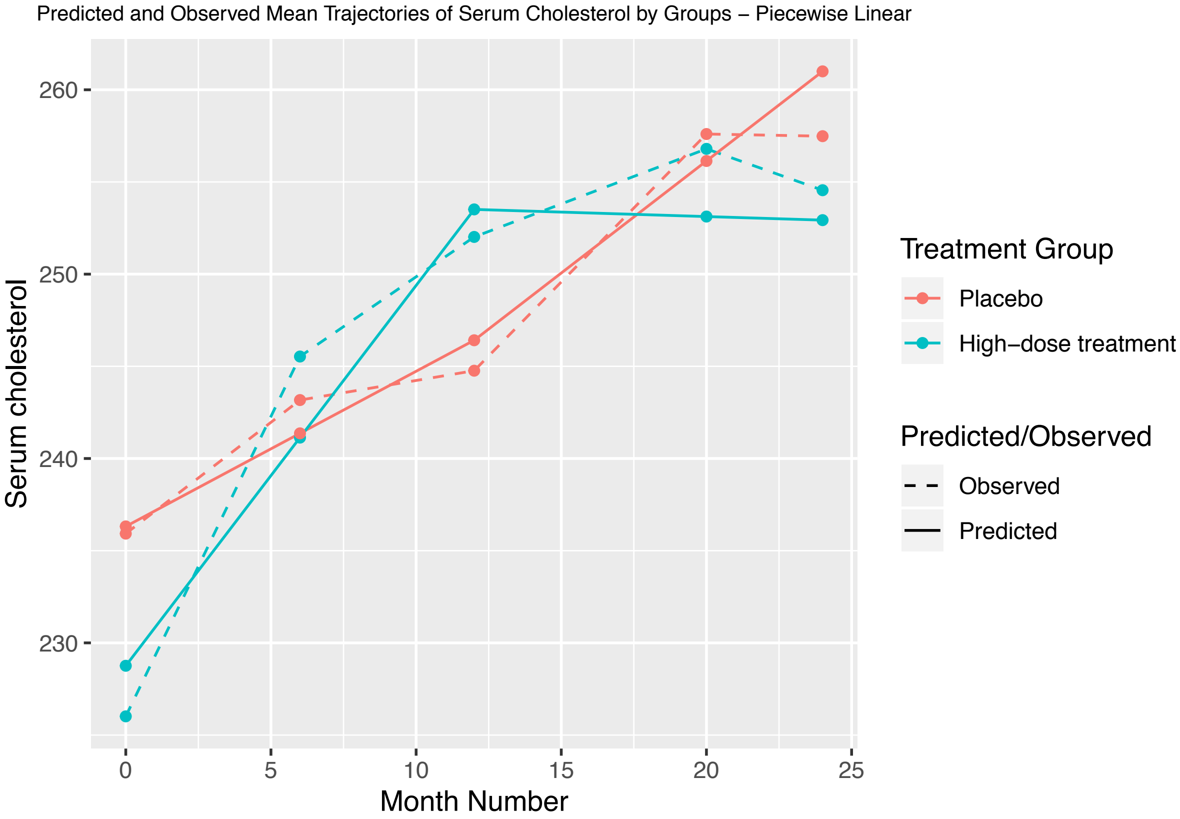

Figure 5: Predicted and Observed Mean Trajectories of Serum Cholesterol by Groups using Piecewise Linear model

### b. Mean in the Treatment (High-dose) Group at month 20

Since $X_1 = 20 > 12$ month, we have

$$\begin{aligned}\mu_{ij}(High-dose, \text{month } 20) &= (\beta_1+\beta_4) - 12(\beta_3+\beta_6) + 20\times(\beta_2+\beta_3+\beta_5+\beta_6)\\ &= 221.20 + 7.56 - 12(-4.60 + 2.49) + 20(3.28 - 4.60 - 1.22 + 2.49)\\ &= 253.08\end{aligned}$$

### j. Plot the observed and predicted from this model mean serum cholesterol levels for each group over time (these should be line plots):

```
# Repeat the above procedure for the piecewise linear model
p.model4 <- unique(data.frame(cgroup=dt.long$cgroup,cmonth=dt.long$cmonth,serum=model4$fitted))
p.model4$group <- "Predicted"
p.model4 <- rbind(p.model4,dt.long_rm_missing_subset)

# Plot predicted and observed mean trajectories by treatment group for the piecewise linear model
p4 <- ggplot(p.model4, aes(x=as.numeric(as.character(cmonth)),y=serum,color=cgroup,linetype=group,group=
p4 + geom_line() + geom_point() +
  labs(title="Predicted and Observed Mean Trajectories of Serum Cholesterol by Groups - Piecewise Linea
       x="Month Number", y="Serum cholesterol", color="Treatment Group", linetype="Predicted/Observed")
  scale_linetype_manual(values=c("dashed","solid"), labels=c("Observed","Predicted")) + theme(plot.titl
```

**v. What are your observations?**

The predicted mean serum cholesterol levels for each group from the mean response model is close to the observed value. It follows the pattern of the observed value, although there are some error. The prediction by the piecewise linear model looks better than the linear model. The prediction of high-dose treatement is under-estimated at time 6, 20, 24.

**Question 5 (10 points): Among the following models: the response profile model, the two parametric time models (linear and quadratic time trends), and the linear piece-wise model with a knot at 12 months – pick the best model for the comparison of the mean levels of serum cholesterol over time:**

Since we haven't build the quadratic time trend model, let us do it now, although we may re-do it use ML(maximum likelihood) to do it again later.

```
#Center month and create a quadratic term of center_month2
dt.long$center_month <- dt.long$month - mean(dt.long$month)
dt.long$center_month2 <- (dt.long$center_month)^2

### Fit a quadratic trend model
model5 <- gls(serum ~ cgroup*center_month+cgroup*center_month2,
              data=dt.long,
              corr=corSymm(form= ~ time | ID),
              weights=varIdent(form= ~ 1 | cmonth),
              na.action = na.omit)
summary(model5)
```

```
## Generalized least squares fit by REML
##   Model: serum ~ cgroup * center_month + cgroup * center_month2
##   Data: dt.long
##        AIC      BIC    logLik
##   4361.722 4447.592 -2159.861
##
## Correlation Structure: General
##  Formula: ~time | ID
##  Parameter estimate(s):
##  Correlation:
##   1     2     3     4
## 2 0.771
## 3 0.732 0.774
## 4 0.738 0.800 0.726
## 5 0.586 0.667 0.677 0.622
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | cmonth
##  Parameter estimates:
##         0         6        12        20        24
## 1.0000000 0.9318753 0.8787859 0.8950787 1.0277396
##
## Coefficients:
##                                              Value Std.Error  t-value
## (Intercept)                              246.27353  6.233253 39.50963
## cgroupHigh-dose treatment                  5.35339  8.048684  0.66513
## center_month                               1.01712  0.219407  4.63578
## center_month2                              0.00513  0.028816  0.17786
## cgroupHigh-dose treatment:center_month    0.13181  0.288809  0.45640
## cgroupHigh-dose treatment:center_month2  -0.09877  0.037841 -2.61017
##                                          p-value
## (Intercept)                               0.0000
## cgroupHigh-dose treatment                 0.5063
## center_month                              0.0000
## center_month2                             0.8589
## cgroupHigh-dose treatment:center_month    0.6483
## cgroupHigh-dose treatment:center_month2   0.0094
##
##  Correlation:
##                                          (Intr) cgrH-t cntr_m cntr_2
## cgroupHigh-dose treatment                -0.774
## center_month                             -0.074  0.058
## center_month2                            -0.238  0.184 -0.157
## cgroupHigh-dose treatment:center_month    0.057 -0.067 -0.760  0.119
## cgroupHigh-dose treatment:center_month2   0.181 -0.239  0.120 -0.761
##                                          cgH-t:_
## cgroupHigh-dose treatment
## center_month
## center_month2
## cgroupHigh-dose treatment:center_month
## cgroupHigh-dose treatment:center_month2  -0.131
##
## Standardized residuals:
##          Min           Q1          Med           Q3          Max
## -2.274885655 -0.678425785 -0.005890717  0.650332884  3.897874298
##
## Residual standard error: 46.7952
## Degrees of freedom: 447 total; 441 residual
```

### a. Conduct appropriate statistical tests

Hint: you should use the Likelihood Ratio test and AIC to compare relevant models, i.e. use these two tests to compare models that are nested within each other and models that are not nested, respectively. The -2log-likelihood and the AIC values should be outputted under the Model Fit Statistics in SAS and R.

We have four models. Note that

$$E(Y_{ij}, \text{mean response}) = \beta_0+\beta_1 group+\beta_2 month6+\beta_3 month12+\beta_4 month20+\beta_5 month24+\beta_6 month6\cdot group+\beta_7 month12\cdot gro$$

$$E(Y_{ij}, \text{linear}) = \beta_0 + \beta_1 time + \beta_2 group + \beta_3 time \cdot group$$

$$E(Y_{ij}, \text{piecewise linear}) = \beta_0+\beta_1 time+\beta_2 (time-t^*)_+ +\beta_3 group+\beta_4 time\cdot group+\beta_5 (time-t^*)_+ \cdot group \text{ where } t^* = 12$$

$$E(Y_{ij}, \text{quadratic}) = \beta_0 + \beta_1 time + \beta_2 time^2 + \beta_3 group + \beta_4 time \cdot group + \beta_5 time^2 \cdot group$$

Therefore, linear model $\subseteq$ quadratic model and linear model $\subseteq$ linear spline model. We can use Likelihood Ratio test to compare. For other pairs of model, we use AIC to compare.
1. compare linear and quadratic

```
m_quadratic <- gls(serum ~ cgroup*center_month+cgroup*center_month2,
                data = dt.long,
                corr = corSymm(form= ~ time | ID),
                weights = varIdent(form= ~ 1 | cmonth),
                na.action = na.omit,
                method='ML')

anova(m_quadratic, m2_full) #m2_full is the linear model
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m_quadratic       1 21 4358.539 4444.693 -2158.269
## m2_full           2 19 4368.149 4446.098 -2165.075 1 vs 2 13.61036  0.0011
```

The p value $0.0011<0.05$ shows that the two models are statistically significant different. We want to maximize the likelihood. Since the sign of the *logLik* column is negative and $2158<2165$, we conclude the quadratic model is better than the linear one. We can use $-2 \times logLik$ to get -2log-likelihood.

2. quadratic and piecewise linear

```
anova(m_quadratic, m4_full) # m4_full is the piecewise linear model
```

```
##               Model df      AIC      BIC    logLik
## m_quadratic       1 21 4358.539 4444.693 -2158.269
## m4_full           2 21 4361.343 4447.496 -2159.671
```

Quadratic model's AIC = 4358.539 which is less than piecewise linear's AIC = 4361.343. Therefore, we conclude that quadratic model is better comparing with the piecewise linear model.

3. quadratic and mean response model

```
anova(m_quadratic, m_full) # m_full is the mean response model
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m_quadratic       1 21 4358.539 4444.693 -2158.269
## m_full            2 25 4363.233 4465.797 -2156.616 1 vs 2 3.306322  0.5079
```

Again, quadratic model has smaller AIC = 4358.539 comparing with mean response model's AIC = 4363.233. Therefore, we conclude that the quadratic model is the best model out of these four models.

```
#Repeat the above procedure for the quadratic model
p.model5 <- unique(data.frame(cgroup=dt.long$cgroup,cmonth=dt.long$cmonth,serum=model5$fitted))
p.model5$group <- "Predicted"
p.model5 <- rbind(p.model5,dt.long_rm_missing_subset)

# Plot predicted and observed mean trajectories by treatment group for the piecewise linear model
p5 <- ggplot(p.model5, aes(x=as.numeric(as.character(cmonth)),y=serum,color=cgroup,linetype=group,group
p5 + geom_line() + geom_point() +
  labs(title="Predicted and Observed Mean Trajectories of Serum Cholesterol by Groups - Quadratic",
       x="Month Number", y="Serum Cholesterol", color="Treatment Group", linetype="Predicted/Observed")
  scale_linetype_manual(values=c("dashed","solid"), labels=c("Observed","Predicted")) + theme(plot.titl
```

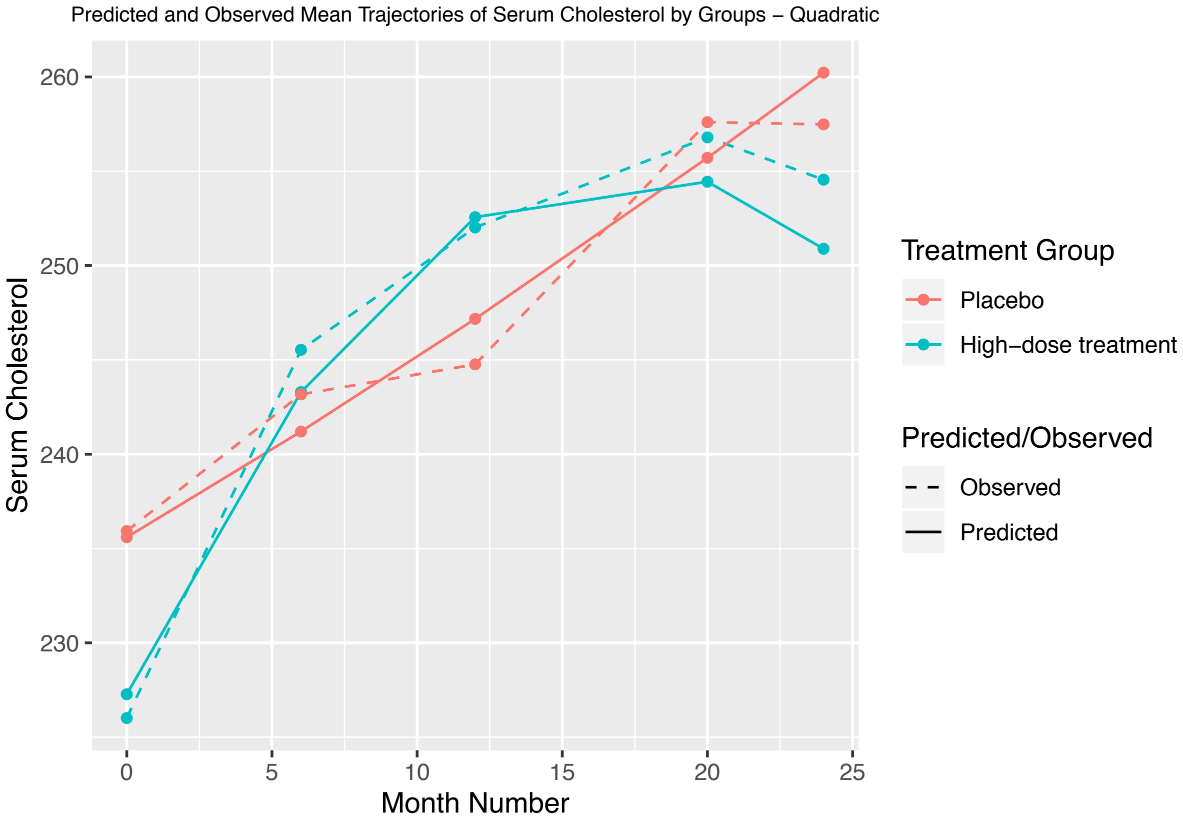

Figure 6: Predicted and Observed Mean Trajectories of Serum Cholesterol by Groups using Quadratic model