BIOST 540 Homework 2

Ivy Zhang, Hantong Hu

4/19/2021

Response

Part A

Question 1

The graphical and numerical summary of the relationship between serum cholesterol levels over time for the two groups have shown below. From the figure and the table, we can see for both groups, generally speaking, the mean serum cholesterol level is increasing over time, except for the time between the month of 20 and the month of 24 in the placebo group. The rate of both groups does not seem to be constant between each time point. At the baseline, it seems the placebo group has a higher mean serum cholesterol level than the high-dose group. At the month of 24 and the month of 20, the two groups tend to have similar mean serum cholesterol levels. At the month of 6 and the month of 12, the high-dose group tends to have a higher mean serum cholesterol level than the placebo group.

It also leads to the increased rate of the high-dose group between the baseline and the month of 12. The placebo group has a higher increase rate between the month of 12 and the month of 24 compared to the high-dose group.

For the correlation among the measurements in different time points, both overall and in the separate groups, the correlation is all positive. In general, the placebo group has a higher correlation value than the high-dose group. Also, in general, measurements are taken in closer time points tend to have higher(or at least similar) correlation values.

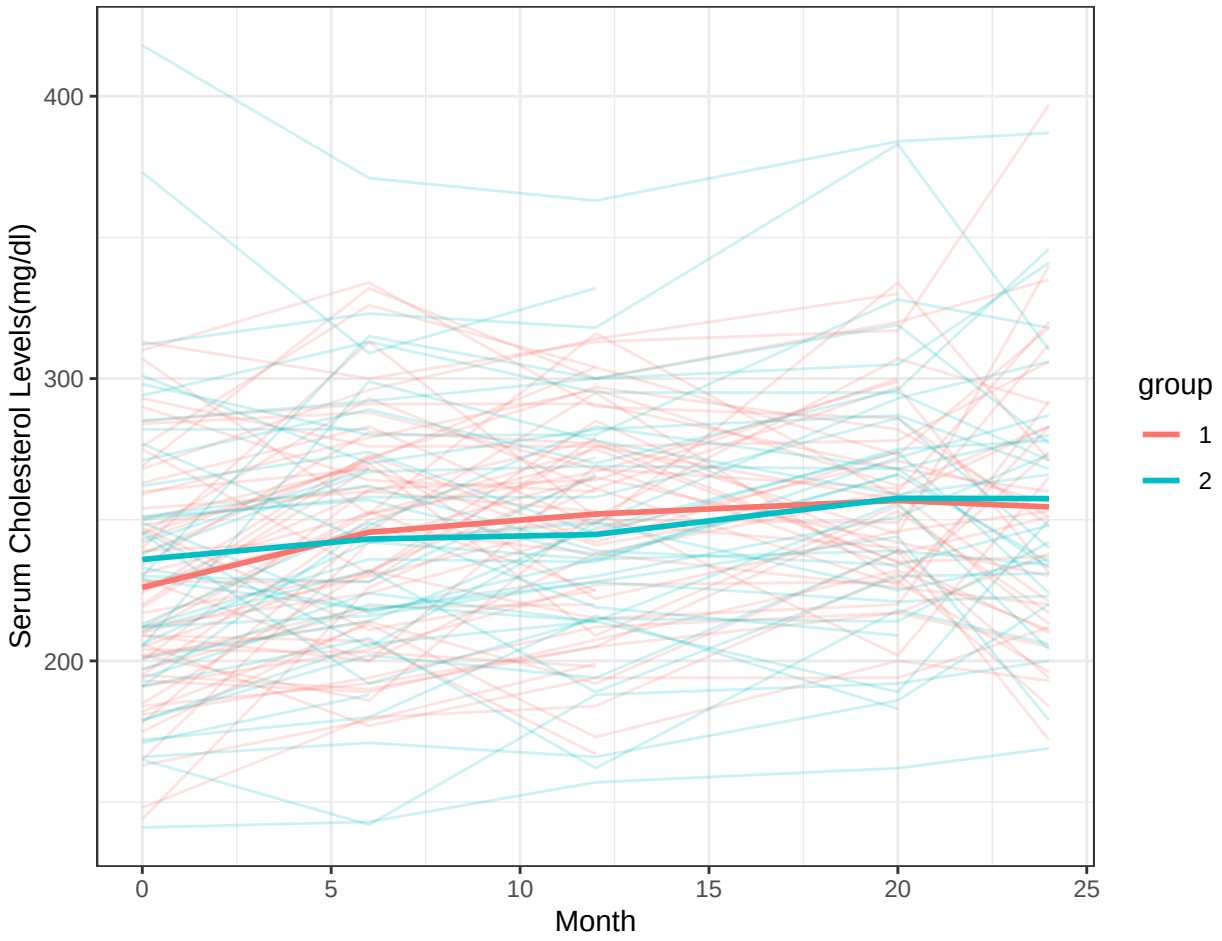

Table 1: High-Dose Summary

baseline	6 months	12 months	20 months	24 months
Min. :144.0	Min. :177.0	Min. :173.0	Min. :194.0	Min. :172.0
1st Qu.:194.0	1st Qu.:221.5	1st Qu.:220.2	1st Qu.:230.2	1st Qu.:216.0
Median :222.5	Median :252.0	Median :252.0	Median :247.5	Median :252.0
Mean :226.8	Mean :249.6	Mean :252.6	Mean :253.1	Mean :256.7
3rd Qu.:255.2	3rd Qu.:276.8	3rd Qu.:286.2	3rd Qu.:271.5	3rd Qu.:285.0
Max. :313.0	Max. :334.0	Max. :316.0	Max. :334.0	Max. :397.0

Table 2: Placebo Summary

baseline	6 months	12 months	20 months	24 months
Min. :141.0	Min. :142.0	Min. :157.0	Min. :162.0	Min. :169.0
1st Qu.:197.5	1st Qu.:211.0	1st Qu.:214.5	1st Qu.:227.5	1st Qu.:227.5
Median :233.0	Median :236.0	Median :245.0	Median :266.0	Median :248.0
Mean :236.6	Mean :243.3	Mean :244.5	Mean :261.9	Mean :257.5
3rd Qu.:266.5	3rd Qu.:277.0	3rd Qu.:279.0	3rd Qu.:290.0	3rd Qu.:279.0
Max. :418.0	Max. :371.0	Max. :363.0	Max. :384.0	Max. :387.0

Table 3: correlation for overall

	y1.resid	y2.resid	y3.resid	y4.resid	y5.resid
y1.resid	1.0000000	0.7641399	0.7479848	0.7580608	0.6056681
y2.resid	0.7641399	1.0000000	0.8070027	0.8215392	0.6945071
y3.resid	0.7479848	0.8070027	1.0000000	0.7414999	0.7035658
y4.resid	0.7580608	0.8215392	0.7414999	1.0000000	0.6496422
y5.resid	0.6056681	0.6945071	0.7035658	0.6496422	1.0000000

Table 4: correlation for high-dose group

	y1.resid	y2.resid	y3.resid	y4.resid	y5.resid
y1.resid	1.0000000	0.7097680	0.6590338	0.6345956	0.4515519
y2.resid	0.7097680	1.0000000	0.6794303	0.7294584	0.5739744
y3.resid	0.6590338	0.6794303	1.0000000	0.5218361	0.6363163
y4.resid	0.6345956	0.7294584	0.5218361	1.0000000	0.5070192
y5.resid	0.4515519	0.5739744	0.6363163	0.5070192	1.0000000

Table 5: correlation for placebo group

	y1.resid	y2.resid	y3.resid	y4.resid	y5.resid
y1.resid	1.0000000	0.8041079	0.8167669	0.8390164	0.7612846
y2.resid	0.8041079	1.0000000	0.9046082	0.8824122	0.8191013
y3.resid	0.8167669	0.9046082	1.0000000	0.8884498	0.7776534
y4.resid	0.8390164	0.8824122	0.8884498	1.0000000	0.7892396
y5.resid	0.7612846	0.8191013	0.7776534	0.7892396	1.0000000

Question2

For this case, I am planning to apply model:

E[mean serum cholesterol$_{ij}|X_i, t_j] = \beta_0 + \beta_1$high-dose$_i + \beta_2 I_{(t_j=2)} + \beta_3 I_{(t_j=3)} + \beta_4 I_{(t_j=4)} + \beta_5 I_{(t_j=5)} + \beta_6$high-dose$_i \times I_{(t_j=2)} + \beta_7$high-dose$_i \times I_{(t_j=3)} + \beta_8$high-dose$_i \times I_{(t_j=4)} + \beta_9$high-dose$_i \times I_{(t_j=5)}$

At here, high-dose$_i$ will equals to 1 if it is an observation from participant$_i$ is from high-dose group and otherwise equals to 0.

$I_{(t_j=2)} = 1$ means the observation is measured at time of 6 months, otherwise equals to 0.

$I_{(t_3=1)} = 1$ means the observation is measured at time of 12 months, otherwise equals to 0.

$I_{(t_j=4)} = 1$ means the observation is measured at time of 20 months, otherwise equals to 0.

$I_{(t_j=5)} = 1$ means the observation is measured at time of 24 months, otherwise equals to 0.

Interpretation:

β_0: The mean serum cholestrol of participants in the placebo group measured at the baseline.

β_1: The difference in mean serum cholestrol of participants between the placebo group and high-dose group at the baseline.

β_2: The difference in mean serum cholestrol of participants between the baseline and the time of month 6 in the placebo group.

β_3: The difference in mean serum cholestrol of partipants between the baseline and the time of month 12 in the placebo group.

β_4: The difference in mean serum cholestrol of partipants between the baseline and the time of month 20 in the placebo group.

β_5: The difference in mean serum cholestrol of participants between the baseline and the time of month 24 in the placebo group.

β_6: The difference in mean serum cholestrol of participants between the baseline and the time of month 6 comparing high-dose group and placebo group.

β_7: The difference in mean serum cholestrol of participants between the baseline and the time of month 12 comparing high-dose and placebo group.

β_8: The difference in mean serum cholestrol of participants between the baseline and the time of month 20 comparing high-dose and placebo group.

β_9: The difference in mean serum cholestrol of participants between the baseline and the time of month 24 comparing high-dose and placebo group.

Question 3

From the summary of our GLS model using Restricted maximum likelihood estimation, it seems that at the month of 6 and at the month of 12, the difference in mean serum cholesterol of participants between the time and baseline of two groups are significantly different at the significance level of 0.05. However, this pattern is not obviously appeared in the month of 20 and in the month of 24.

Overall, we cannot reject the null hypothesis that the patterns of change over time do not differ between the two groups at the 5% significance level($p = 0.099$) using the ANOVA test.

Question 4

From the summary of our GLS model using Restricted maximum likelihood estimation, we estimate that for individuals in the placebo group and differing in one month in measuring time, the group that is measured later has mean serum cholesterol that is 1.020 mg/dL higher. We also estimate that for individuals in the treatment group and differing in one month in measuring time, the group that is measured later has mean serum cholesterol that is 1.207 mg/dL higher. We find no statistically significant difference in the patterns of change in the time between the two groups ($p = 0.7366$) at the significance level of 0.05.

Question 5

Although both the conclusion of the two models is there is no statistically significant difference in the rate of increase in mean serum cholesterol levels between the two groups. The categorical time model shows a relatively low p-value than the linear time model. To the categorical model, if there is no difference in the rate of increase in mean serum cholesterol levels between the two groups, it is less possible to appear the situation or more extreme situation of this data compared to the linear time model.

Treating time categorically:

Advantage: For each point, we can compare the difference between the four-time points and the baseline in these two groups independently from other time points.

Disadvantage: The measurements are not made at equal intervals of time. We only can estimate the difference in these four-time points. We cannot estimate the difference between other time points and the baseline comparing the two groups.

Treating time linearly:

Advantage: We can estimate the difference between the time point and the baseline by comparing two groups at any time point using our GLS model.

Disadvantage: We need to assume at every time point, the rate of increase in mean serum cholesterol levels is the same for both the placebo group and treatment group.

Part B

Question1

$$E(distance_{ij}|X_i, age_j) = \beta_0 + \beta_1 * female_i + \beta_2 * age_j + \beta_3 * female_i \times age_j$$

β_0 is the mean distance for male at age=0, β_1 is the difference in mean distance between male and female at age=0, β_2 is the difference in mean distance for 1mm increase in distance, and β_3 is the difference in difference in mean distance for 1mm increase in distance comparing male and female group.

Question 2

Table 6: OLS, model-based standard errors (homoscedasticity)

	Estimate	Std. Error
(Intercept)	16.3406250	1.4162242
SexFemale	1.0321023	2.2187969
age	0.7843750	0.1261673
SexFemale:age	-0.3048295	0.1976661

Table 7: GLS, unstructured/symmetric correlation matrix, heteroscedasticity, REML

	Estimate	Std.Error
(Intercept)	15.8422827	0.9723039
relevel(Sex, “Male”)Female	1.5830863	1.5233074
age	0.8268037	0.0822177
relevel(Sex, “Male”)Female:age	-0.3504390	0.1288104

Table 8: GLS, exchangeable/compound symmetric correlation matrix, homoscedasticity, REML

	Estimate	Std.Error
(Intercept)	16.3406250	0.9813122
relevel(Sex, “Male”)Female	1.0321023	1.5374208
age	0.7843750	0.0775011
relevel(Sex, “Male”)Female:age	-0.3048295	0.1214209

Table 9: LMM, random intercepts, REML

	Estimate	Std.Error
(Intercept)	16.3406250	0.9813122
SexFemale	1.0321023	1.5374208
age	0.7843750	0.0775011
SexFemale:age	-0.3048295	0.1214209

Table 10: LMM, random intercepts + slopes (correlated), REML

	Estimate	Std.Error
(Intercept)	16.3406250	1.0185318
SexFemale	1.0321023	1.5957326
age	0.7843750	0.0859995
SexFemale:age	-0.3048295	0.1347353

Question 3

a) **OLS (mod1)** assumes independent observations, so it assumes no correlation structure, which is displayed as only diagnol having non-zero values and others 0; covariance matrices would have (same) variance on the diagnol and others 0.

GLS (unstructered corr) (mod2) assumes unstructured correlation matrix. Since it also assumes heteroscedasticity, so it will have different values for every cell (except those holding for symmetry) in the covariance matrix.

GLS (exchangeable corr) (mod3) assumes exchangeable correlation matrix. Since it also assumes homoscedasticity, so it will have the same values for every cell ($\rho * \sigma^2$) except the diagonol, and the values for the diagonol are the same (σ^2) too, in the covariance matrix.

LMM (random intercept) (mod4) assumes exchangeable correlation matrix. The covariance matrix will have the same values for every cell (G_{11}) except the diagonol, and the values for the diagonol are the same ($G_{11} + \sigma^2$) too.

LMM (random intercept&slope) (mod5) assumes unstructured correlation matrix, and it will have different values for every cell (except those holding for symmetry) in the covariance matrix. The diagnol values will be $G_{11} + G_{22}t_{ij}^2 + 2G_{12}t_{ij} + sigma^2$, the other cells will be the covariance value between Y_{ij}andY_{ik} will be $G_{11} + G_{12}(t_{ij} + t_{ik}) + G_{22}t_{ij}t_{ik}$.

b) For the point estimates, all except the **GLS unstructured corr** displayed the same results. For the SE, **OLS** has the highest SE, and other 4 models have similar SE with **LMM (random intercept&slope)** having slightly higher values of SE.

c) For the point estimates and standard errors that OLS provides, the point estimates are unbiased and valid but the model-based standard errors are not valid.

Code Appendix

```
### Setting up the packages, options we'll need:
library(nlme)
library(dplyr)
library(reshape2)
library(ggplot2)
library(joineR)
library(MASS)
library(magrittr)
library(knitr)
library(uwIntroStats)
library(nlme)
data(Orthodont)
### -------------------------------------------------------------
### Reading in the data for part A.
cholesterol <- read.csv("~/Desktop/R hw/cholesterol.csv")
cholesterol = cholesterol[,-1]
cholesterol_long = melt(cholesterol, id=c("id","group"))
cholesterol_long$month <- (as.numeric(gsub("y","",cholesterol_long$variable))-1)*6
for(i in 1: nrow(cholesterol_long)){
  if(cholesterol_long[i,"month"] == 18){
    cholesterol_long[i,"month"] = 20
  }
}
cholesterol_long$time <- as.numeric(as.factor(cholesterol_long$month))
cholesterol_long$group = factor(cholesterol_long$group)
cholesterol = na.omit(cholesterol)
cholesterol_long = na.omit(cholesterol_long)
### Reading in the data for part B
ortho_wide <- reshape(Orthodont,
                      direction="wide",
                      idvar = c("Subject", "Sex"),
                      timevar="age")
Orthodont$time <- as.numeric(as.factor(Orthodont$age))
###-------------------------------------------------------------
###Part A
### -------------------------------------------------------------
### Q1
p <- ggplot(data = cholesterol_long, aes(x = month, y = value, group = id, col = group))
p + geom_line(alpha=0.2) +
  geom_line(data = cholesterol_long %>% group_by(group, month) %>% summarise(value=mean(value)),
            aes(x = month, y = value, group = group), size=1) +
  theme_bw() + ylab("Serum Cholesterol Levels(mg/dl)")+xlab("Month")
stat = by(cholesterol[,-c(1,2)], INDICES = cholesterol$group, FUN = summary)
knitr::kable(stat[[1]], col.names = c("baseline","6 months","12 months","20 months","24 months"),
             align = "rrrrrr",
             caption = "High-Dose Summary")
knitr::kable(stat[[2]],col.names = c("baseline","6 months","12 months","20 months","24 months"),
             align = "rrrrrr",
             caption = "Placebo Summary")

chol_means <- cholesterol %>% group_by(group) %>%
```

```
  summarise(y5mean = mean(y5), y1mean=mean(y1),
            y4mean=mean(y4), y2mean=mean(y2), y3mean=mean(y3))
cholesterol <- merge(cholesterol, chol_means, by="group")
cholesterol$y2.resid <- cholesterol$y2 - cholesterol$y2mean
cholesterol$y1.resid <- cholesterol$y1 - cholesterol$y1mean
cholesterol$y4.resid <- cholesterol$y4 - cholesterol$y4mean
cholesterol$y3.resid <- cholesterol$y3 - cholesterol$y3mean
cholesterol$y5.resid = cholesterol$y5 - cholesterol$y5mean

kable(cor(cholesterol[,c("y1.resid", "y2.resid", "y3.resid", "y4.resid","y5.resid")]),
      caption = "correlation for overall")
corr_summary = by(cholesterol[,c("y1.resid", "y2.resid", "y3.resid", "y4.resid","y5.resid")],INDICES = c
kable(corr_summary[[1]], caption = "correlation for high-dose group")
kable(corr_summary[[2]], caption = "correlation for placebo group")
### ------------------------------------------------------------
### Q3
c.reml1 =  gls(value ~ relevel(group,2)*as.factor(month),
               data=cholesterol_long,
               method="REML",
               correlation=corSymm(form = ~time | id),
               weights=varIdent(form= ~1 | month))
summary(c.reml1)
anova(c.reml1)
### ------------------------------------------------------------
### Q4
c.reml2 =  gls(value ~ relevel(group,2)*month,
               data=cholesterol_long,
               method="REML",
               correlation=corSymm(form = ~time | id),
               weights=varIdent(form= ~1 | month))
summary(c.reml2)
anova(c.reml2)
### ------------------------------------------------------------
### QB1
library(nlme)
data(Orthodont)
### ------------------------------------------------------------
### QB2
# table <- matrix(NA, nrow=5, ncol = 4)

# OLS, model-based standard errors (homoscedasticity)
# mod1 <- lm(distance~age, data = Orthodont)
# summary(mod1)
mod1a <- lm(distance~Sex*age, data = Orthodont)
mod1.dat <- as.data.frame(coef(summary(mod1a))[,c(1,2)])

# GLS, unstructured/symmetric correlation matrix, heteroscedasticity, REML
Orthodont$time <- as.numeric(as.factor(Orthodont$age))
mod2 <- gls(distance ~ relevel(Sex,"Male")*age,
            data = Orthodont, method="REML",
            correlation=corSymm(form = ~time | Subject),
            weights=varIdent(form= ~1 | age))
mod2.dat <- as.data.frame(coef(summary(mod2)))
```

```r
mod2.dat$Estimate <- mod2.dat$Value
mod2.dat <- mod2.dat[,c(5,2)]

# GLS, exchangeable/compound symmetric correlation matrix, homoscedasticity, REML
mod3 <- gls(distance ~ relevel(Sex,"Male")*age,
            data = Orthodont, method="REML",
            correlation=corCompSymm(form = ~time | Subject))
mod3.dat <- as.data.frame(coef(summary(mod3)))
mod3.dat$Estimate <- mod3.dat$Value
mod3.dat <- mod3.dat[,c(5,2)]

# LMM, random intercepts, REML
mod4 <- lme(distance ~ Sex*age,
             method = "REML", data = Orthodont ,
             random = reStruct( ~ 1 | Subject, pdClass="pdDiag", REML=T))
mod4.dat <- as.data.frame(coef(summary(mod4)))
mod4.dat$Estimate <- mod4.dat$Value
mod4.dat <- mod4.dat[,c(6,2)]

# LMM, random intercepts + slopes (correlated), REML
mod5 <- lme(distance ~ Sex*age,
             method = "REML", data = Orthodont ,
             random = reStruct( ~ 1+age | Subject, pdClass="pdSymm", REML=T))
mod5.dat <- as.data.frame(coef(summary(mod5)))
mod5.dat$Estimate <- mod5.dat$Value
mod5.dat <- mod5.dat[,c(6,2)]

ls <- list(df1 = mod1.dat, df2 = mod2.dat,
           df3 = mod3.dat, df4=mod4.dat, df5=mod5.dat)
knitr::kable(ls[[1]], caption = "OLS, model-based standard errors (homoscedasticity)")
knitr::kable(ls[[2]], caption = "GLS, unstructured/symmetric correlation matrix, heteroscedasticity, RE
knitr::kable(ls[[3]], caption = "GLS, exchangeable/compound symmetric correlation matrix, homoscedastic
knitr::kable(ls[[4]], caption = "LMM, random intercepts, REML")
knitr::kable(ls[[5]], caption = "LMM, random intercepts + slopes (correlated), REML")

### ------------------------------------------------------------
### QB3
```