Homework 2

Meixin Zhang

Part A

1. Conduct an exploratory data analysis by using graphical methods and numerical summary statistics to describe (1) the relationship between serum cholesterol levels over time for the two groups and (2) the correlation between measurements. Write a paragraph commenting on your observations from the exploratory data analysis.



(1) The numerical summary statistics to describe the relationship between serum cholesterol levels over time for the two groups

## cholesterol$group: 1  
## y1 y2 y3 y4   
## Min. :144.0 Min. :177.0 Min. :167.0 Min. :194.0   
## 1st Qu.:200.2 1st Qu.:214.0 1st Qu.:223.5 1st Qu.:232.8   
## Median :222.0 Median :248.5 Median :260.0 Median :254.0   
## Mean :226.0 Mean :245.5 Mean :252.0 Mean :256.8   
## 3rd Qu.:249.5 3rd Qu.:271.8 3rd Qu.:277.5 3rd Qu.:275.0   
## Max. :313.0 Max. :334.0 Max. :316.0 Max. :334.0   
## NA's :7 NA's :18   
## y5   
## Min. :172.0   
## 1st Qu.:214.0   
## Median :251.0   
## Mean :254.6   
## 3rd Qu.:283.0   
## Max. :397.0   
## NA's :24   
## --------------------------------------------------------   
## cholesterol$group: 2  
## y1 y2 y3 y4   
## Min. :141.0 Min. :142.0 Min. :157.0 Min. :162.0   
## 1st Qu.:197.0 1st Qu.:214.0 1st Qu.:215.2 1st Qu.:223.0   
## Median :230.0 Median :242.0 Median :238.0 Median :259.0   
## Mean :235.9 Mean :243.2 Mean :244.8 Mean :257.6   
## 3rd Qu.:262.0 3rd Qu.:274.0 3rd Qu.:276.0 3rd Qu.:286.5   
## Max. :418.0 Max. :371.0 Max. :363.0 Max. :384.0   
## NA's :3 NA's :6   
## y5   
## Min. :169.0   
## 1st Qu.:227.5   
## Median :248.0   
## Mean :257.5   
## 3rd Qu.:279.0   
## Max. :387.0   
## NA's :10

(2) The correlation between measurements

## y0.resid y6.resid y12.resid y20.resid y24.resid  
## y0.resid 1.0000000 0.7634717 0.7479556 0.7547128 0.6055315  
## y6.resid 0.7634717 1.0000000 0.8066409 0.8132653 0.6946468  
## y12.resid 0.7479556 0.8066409 1.0000000 0.7373536 0.7035717  
## y20.resid 0.7547128 0.8132653 0.7373536 1.0000000 0.6446018  
## y24.resid 0.6055315 0.6946468 0.7035717 0.6446018 1.0000000

## choles\_complete$group: 1  
## y0.resid y6.resid y12.resid y20.resid y24.resid  
## y0.resid 1.0000000 0.7097680 0.6590338 0.6345956 0.4515519  
## y6.resid 0.7097680 1.0000000 0.6794303 0.7294584 0.5739744  
## y12.resid 0.6590338 0.6794303 1.0000000 0.5218361 0.6363163  
## y20.resid 0.6345956 0.7294584 0.5218361 1.0000000 0.5070192  
## y24.resid 0.4515519 0.5739744 0.6363163 0.5070192 1.0000000  
## --------------------------------------------------------   
## choles\_complete$group: 2  
## y0.resid y6.resid y12.resid y20.resid y24.resid  
## y0.resid 1.0000000 0.8041079 0.8167669 0.8390164 0.7612846  
## y6.resid 0.8041079 1.0000000 0.9046082 0.8824122 0.8191013  
## y12.resid 0.8167669 0.9046082 1.0000000 0.8884498 0.7776534  
## y20.resid 0.8390164 0.8824122 0.8884498 1.0000000 0.7892396  
## y24.resid 0.7612846 0.8191013 0.7776534 0.7892396 1.0000000

From the figure and the numerical summary of the mean cholesterol over time, we can know that in general, the cholesterol increase over time in both treatment and control groups. And in treatment group, the baseline cholesterol level is lower than the baseline cholesterol in control group, but increased at a higher rate than control group at early stages (from baseline to week12). After week12, the increase rate of cholesterol level became lower in treatment group than in the control group. From week 20 to week 24, the cholesterol level even dropped in treatment group, whereas cholesterol keep constant in control group.

And for the correlation of measurement, the correlation is larger in the control group than that in the treatment group. And in general, the correlation decrease as the time lag became larger in both treatment and control group.

1. With baseline (month 0) and the placebo group (group 2) as the reference group, write out the regression model for mean serum cholesterol that corresponds to the analysis of response profiles. Provide interpretations for each coefficient.

E(Yij|Xi,tj)=β0+β1\*treatment+ β2\*I6 + β3\*I12 + β4\*I20 + β5\* I24+ β6\*treatmenti\* I6 + β7\*treatmenti\* I12 + β8\*treatmenti\* I20 + β9\*treatmenti\* I24

β0: the mean serum cholesterol at baseline for placebo group

β1: the mean serum cholesterol at baseline for treatment group

β2: the difference in mean serum cholesterol between 6 month and baseline for placebo group

β3: the difference in mean serum cholesterol between 12 month and baseline for placebo group

β4: the difference in mean serum cholesterol between 20 month and baseline for placebo group

β5: the difference in mean serum cholesterol between 24 month and baseline for placebo group

β6: the difference in difference in mean serum cholesterol between 6 month and baseline between treatment group and placebo group

β7: the difference in difference in mean serum cholesterol between 12 month and baseline between treatment group and placebo group

β8: the difference in difference in mean serum cholesterol between 20 month and baseline between treatment group and placebo group

β9: the difference in difference in mean serum cholesterol between 24 month and baseline between treatment group and placebo group

1. Using GLS, conduct an analysis of response profiles (using all available measurements) with an unstructured covariance matrix and allowing for heteroscedasticity. Determine whether the patterns of change over time differ between the two groups. What do you conclude at the 5% significance level?

E(Yij|Xi,tj)=β0+β1\*treatmenti + β2\*I(tj=6) + β3\*I(tj=12) + β4\*I(tj=20) + β5\* I(tj=24) + β6\*treatmenti\* I(tj=6) + β7\*treatmenti\* I(tj=12) + β8\*treatmenti\* I(tj=20) + β9\*treatmenti\* I(tj=24)

H0: β6=β7=β8=β9=0

|  | **Value**  <chr> | **Std.Error**  <chr> | **t-value**  <chr> | **p-value**  <chr> |
| --- | --- | --- | --- | --- |
| (Intercept) | 235.92683 | 7.302782 | 32.30643 | 0.0000 |
| groupTreatment | -9.91070 | 9.412632 | -1.05291 | 0.2930 |
| as.factor(time)2 | 7.24390 | 4.805468 | 1.50743 | 0.1324 |
| as.factor(time)3 | 8.84832 | 5.207268 | 1.69922 | 0.0900 |
| as.factor(time)4 | 23.10279 | 5.297447 | 4.36112 | 0.0000 |
| as.factor(time)5 | 21.12377 | 7.369940 | 2.86621 | 0.0044 |
| groupTreatment:as.factor(time)2 | 12.27223 | 6.193818 | 1.98137 | 0.0482 |
| groupTreatment:as.factor(time)3 | 16.41754 | 6.743418 | 2.43460 | 0.0153 |
| groupTreatment:as.factor(time)4 | 4.97699 | 6.982010 | 0.71283 | 0.4763 |
| groupTreatment:as.factor(time)5 | 6.90311 | 9.791182 | 0.70503 | 0.4812 |

|  | **numDF**  <int> | **F-value**  <chr> | **p-value**  <chr> |
| --- | --- | --- | --- |
| (Intercept) | 1 | 4280.329 | <.0001 |
| group | 1 | 0.006 | 0.9369 |
| as.factor(time) | 4 | 16.471 | <.0001 |
| group:as.factor(time) | 4 | 1.965 | 0.0988 |

Based on the output of GLS model, we can a statistically significant difference between β6 and 0, and β7 and 0. But this differences are not statistically significant for β8 and β9. And the output of Wald test testing the hypothesis demonstrates that we have confidence to reject the null hypothesis that the patterns of change over time don’t differ between the two groups (p-value = 0.099).

1. Now consider a parametric formulation of the mean profiles with only a linear trend in time. Fit the model. What is the estimated rate of increase in mean cholesterol levels for the treatment group? What is the estimated rate of increase in mean serum cholesterol levels for the placebo group? At the 5% significance level, are the patterns of change the same in the two groups?

E(Yij|Xi,tj) = β0+β1\*treatmenti + β2\*c(month) + β3\*treatmenti\*month

|  | **Value**  <chr> | **Std.Error**  <chr> | **t-value**  <chr> | **p-value**  <chr> |
| --- | --- | --- | --- | --- |
| (Intercept) | 235.28055 | 6.779242 | 34.70603 | 0.0000 |
| groupTreatment | -0.86560 | 8.750192 | -0.09892 | 0.9212 |
| month | 1.01972 | 0.218920 | 4.65794 | 0.0000 |
| groupTreatment:month | 0.09761 | 0.290038 | 0.33654 | 0.7366 |

Based on the output, we can see that the estimated rate of increase in mean serum cholesterol for the treatment group is 1.12 mg/dL(1.01972+0.09761) per month. And the estimated rate of increase in mean serum cholesterol for the placebo group is 1.02 mg/dL. At the 5% significance level, the patterns of change are the same in the two groups (p-value = 0.74).

1. Compare the models and conclusions from questions 3 and 4. Discuss the advantages and disadvantages of treating time categorically (i.e. analysis of response profiles) and treating time linearly.

Advantage of treating time categorically:

more flexible;

Allow different outcome difference between groups at different time point;

enable us to compare multiple time point-pairs.

Disadvantage of treating time categorically:

there are too many parameters in the model.

Advantage of treating time linearly:

the model is more simple;

easy to interpret and understand.

Disadvantage of treating time linearly:

not flexible;

Assuming constant change over time

Part B

We are interested in answering the question: is there a difference in dental growth rates between males and females?

1. Write out the regression model with dental length as the outcome to address this question, assuming a linear model for time. What are the interpretations for each of the parameters?

E(Yij|Xi,tj) = β0+β1\* c(age) + β2\* Sexi + β3\*Sexi\*age

β0: the dental distance for male at 0 years old. But because we don’t have any data for dental distance at 0 years old, this parameter has no practical meanings.

β1: the average annually dental growth for male.

β2: the difference in dental distance for female at 0 years old, when comparing with male. Still, this parameter has no meanings.

β3: the difference of average annually dental growth between female and male at same age.

1. Create a table with estimates and standard errors from fitting each of the following models for all the parameters in the mean model:

• OLS, model-based standard errors (homoscedasticity)

• GLS, unstructured/symmetric correlation matrix, heteroscedasticity, REML

• GLS, exchangeable/compound symmetric correlation matrix, homoscedasticity, REML

• LMM, random intercepts, REML

• LMM, random intercepts + slopes (correlated), REML

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Parameters | Estimates | SE |
| OLS, model-based standard errors (homoscedasticity) | Intercept | 21.18 | 0.69 |
| SexMale | 1.69 | 0.90 |
| Age(10) | 1.05 | 0.98 |
| Age(12) | 1.91 | 0.98 |
| Age(14) | 2.91 | 0.98 |
| SexMale:as.factor(age)10 | -0.11 | 1.27 |
| SexMale:as.factor(age)12 | 0.93 | 1.27 |
| SexMale:as.factor(age)14 | 1.68 | 1.27 |
| LMM, random intercepts, REML | Intercept | 21.18 | 069 |
| SexMale | 1.69 | 0.90 |
| Age(10) | 1.05 | 0.60 |
| Age(12) | 1.91 | 0.60 |
| Age(14) | 2.91 | 0.60 |
| SexMale:as.factor(age)10 | -0.11 | 0.78 |
| SexMale:as.factor(age)12 | 0.93 | 0.78 |
| SexMale:as.factor(age)14 | 1.68 | 0.78 |
| LMM, random intercepts + slopes (correlated), REML | Intercept | 21.18 | 0.67 |
| SexMale | 1.69 | 0.87 |
| Age(10) | 1.05 | 0.58 |
| Age(12) | 1.91 | 0.61 |
| Age(14) | 2.91 | 0.65 |
| SexMale:as.factor(age)10 | -0.11 | 0.75 |
| SexMale:as.factor(age)12 | 0.93 | 0.79 |
| SexMale:as.factor(age)14 | 1.68 | 0.84 |
| GLS, unstructured/symmetric correlation matrix, heteroscedasticity, REML | Intercept | 21.18 | 0.70 |
| SexMale | 1.69 | 0.91 |
| Age(10) | 1.05 | 0.62 |
| Age(12) | 1.91 | 0.61 |
| Age(14) | 2.91 | 0.67 |
| SexMale:as.factor(age)10 | -0.11 | 0.80 |
| SexMale:as.factor(age)12 | 0.93 | 0.79 |
| SexMale:as.factor(age)14 | 1.68 | 0.87 |
| GLS, exchangeable/compound symmetric correlation matrix, homoscedasticity, REML | Intercept | 21.18 | 069 |
| SexMale | 1.69 | 0.90 |
| Age(10) | 1.05 | 0.60 |
| Age(12) | 1.91 | 0.60 |
| Age(14) | 2.91 | 0.60 |
| SexMale:as.factor(age)10 | -0.11 | 0.78 |
| SexMale:as.factor(age)12 | 0.93 | 0.78 |
| SexMale:as.factor(age)14 | 1.68 | 0.78 |

1. Comment on each of the models and the results.

OLS,model-based standard errors (homoscedasticity)

The point estimates are unbiased and are the same as other models. But the standard error are not valid, because when using OLS, we assume the data are independent. And the standard errors in this model are larger than the SE in other models.

GLS, unstructured/symmetric correlation matrix, heteroscedasticity, REML

The point estimates are unbiased. We assume different correlations in the matrix, so the standard errors are all different. This model is more flexible than when we assume exchangeable correlation.

GLS, exchangeable/compound symmetric correlation matrix, homoscedasticity, REML

The point estimates are unbiased. We assume same correlations in the matrix, so the standard errors are the same for all time point(when tima lags are not same, the correlations are still same). The standard errors may also be valid because from GLS model assuming unstructured/symmetric correlation matrix, we can tell that the standard errors for different tima lag don’t differ too much.

LMM, random intercepts, REML

The results are exactly same as GLS model assuming exchangeable correlation matrix.

LMM, random intercepts + slopes (correlated), REML

The standard errors increase as time lag increases. This model make more assumptions than LMM only assuming random intercepts. It also assumes the intercepts and slopes are correlated, which means as the intercepts increase, the slopes increase/decrease.

Appendix:

## Part A

rm(list=ls())  
library(reshape2)

library(ggplot2)  
library(dplyr)

library(nlme)

cholesterol <- read.csv("~/Documents/MS-2nd-year/Spring/BIOST540/HW/cholesterol.csv")

## Question 1

cholesterol<-cholesterol[,c("group","id","y1","y2","y3","y4","y5")]  
choles\_long <- melt(cholesterol, id=c("id","group"))  
names(choles\_long)<-c("id","group","time","value")  
choles\_long$time <- as.numeric(gsub("y","",choles\_long$time))  
choles\_long$group[choles\_long$group==1]<-"Treatment"  
choles\_long$group[choles\_long$group==2]<-"Control"  
choles\_means <- aggregate(value ~ time + group, data = choles\_long, FUN=mean)  
p <- ggplot(data = choles\_means, aes(x = time, y = value, group = group,   
 col = group))  
p + geom\_line() + scale\_color\_manual(values=c('blue','red')) +   
 geom\_point() + theme\_bw() + ylab("Serum cholesterol (mg/dL)") + xlab("Time points")

## summary statistics

by(cholesterol[,-c(1,2)], INDICES = cholesterol$group, FUN=summary)

## Estimated correlation matrix after subtracting out means by group and month

choles\_mean <- cholesterol %>% group\_by(group) %>%   
 summarise(y0mean = mean(y1), y6mean=mean(y2),   
 y12mean=mean(y3,na.rm = TRUE), y20mean=mean(y4,na.rm = TRUE),  
 y24mean=mean(y5,na.rm = TRUE))   
cholesterol <- merge(cholesterol, choles\_mean, by="group")  
cholesterol$y0.resid <- cholesterol$y1 - cholesterol$y0mean  
cholesterol$y6.resid <- cholesterol$y2 - cholesterol$y6mean  
cholesterol$y12.resid <- cholesterol$y3 - cholesterol$y12mean  
cholesterol$y20.resid <- cholesterol$y4 - cholesterol$y20mean  
cholesterol$y24.resid <- cholesterol$y5 - cholesterol$y24mean  
choles\_complete<-cholesterol[rowSums(is.na(cholesterol))==0,]  
cor(choles\_complete[,c("y0.resid", "y6.resid", "y12.resid", "y20.resid","y24.resid")])

## Estimated correlation matrix by treatment groups

by(choles\_complete[,c("y0.resid", "y6.resid", "y12.resid", "y20.resid", "y24.resid")],   
 INDICES = choles\_complete$group, FUN=cor)

## Question 3

m.reml<-gls(value~group\*as.factor(time),  
 data=choles\_long[rowSums(is.na(choles\_long))==0,],  
 method="REML",  
 correlation=corSymm(form = ~time | id),  
 weights=varIdent(form= ~1 | time))  
summary(m.reml)

## Question 4

choles\_long$month[choles\_long$time==1]<-0  
choles\_long$month[choles\_long$time==2]<-6  
choles\_long$month[choles\_long$time==3]<-12  
choles\_long$month[choles\_long$time==4]<-20  
choles\_long$month[choles\_long$time==5]<-24  
m.reml2<-gls(value~group\*month,  
 data=choles\_long[rowSums(is.na(choles\_long))==0,],  
 method="REML",  
 correlation=corSymm(form = ~time | id),  
 weights=varIdent(form= ~1 | month))  
summary(m.reml2)

## Part B

data(Orthodont)  
Orthodont$id<-cumsum(!duplicated(Orthodont$Subject))

## Question 2

## OLS

lm<-lm(distance ~ as.factor(age)\*Sex,   
 data = Orthodont)  
summary(lm)

## GLS, unstructured/symmetric correlation matrix, heteroscedasticity, REML

Orthodont$time<-ifelse(Orthodont$age==8,1,  
 ifelse(Orthodont$age==10,2,  
 ifelse(Orthodont$age==12,3,4)))  
reml\_2<-gls(distance ~ Sex\*as.factor(age),  
 data = Orthodont,  
 method = "REML",  
 correlation = corSymm(form = ~time | Subject),  
 weights = varIdent(form = ~1 | as.factor(age)))  
summary(reml\_2)

## GLS, exchangeable/compound symmetric correlation matrix, homoscedasticity, REML

reml\_3<-gls(distance ~ Sex\*as.factor(age),   
 data=Orthodont,   
 method="REML",  
 correlation=corCompSymm(form = ~time | id))  
summary(reml\_3)

## LMM, random intercepts, REML

lmm1 <- lme( distance ~ Sex\*as.factor(age),   
 method = "REML", data = Orthodont,   
 random = reStruct( ~ 1 | id, pdClass="pdDiag", REML=F))  
summary(lmm1)

## LMM, random intercepts + slopes (correlated), REML

lmm2 <- lme( distance ~ Sex\*as.factor(age),   
 method = "REML", data = Orthodont,   
 random = reStruct( ~ 1 + age | id, pdClass="pdSymm", REML=F))  
summary(lmm2)