ST437/537 – HW #06

Arnab Maity

Due date: April 09, 2019

Instructions

Please follow the instructions below when you prepare and submit your assignment.

- **Include a cover-page** with your homework. It should contain

i. Full name,
ii. Course#: ST 437/537 and
iii. HW-#
iv. Submission date

- Assignments should be submitted in class on the date specified ("due date").
- Neatly typed or hand-written solution on standard letter-size papers (stapled on the top-left corner) should be submitted. **All R code/output should be well commented, with relevant outputs highlighted.**
- **Always staple (upper left corner) your homework before coming to class. Ten percent points will be deducted otherwise.**
- When you solve a particular problem, do not only give the final answer. Instead **show all your work** and the steps you used (with proper explanation) to arrive at your answer to get full credit.
- **DO NOT** give printouts of whole dataset or matrices. Present only the relevant output when answering a question.

Problems

Solve the following problems. You may use `R` for these problems unless I specifically instruct otherwise.

DO NOT give printouts of whole dataset or matrices. Present only the relevant output/graphs when answering a question.

Problem 1 (35 points)

In the [study of dental growth] (../data/dental.txt) (Potthoff and Roy, 1964), 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. Refer to Example A in the lecture on [Models for mean and covariance] (../Lecture08_LDA_Modeling_and_Estimation)

(i) Assume a linear model for the mean response for each group. Fit the following models for the covariance:

1. unstructured covariance
2. compound symmetry
3. heterogeneous compound symmetry

4. autoregressive
5. heterogeneous autoregressive

Choose a model for the covariance that adequately fits the data (you may use AIC/BIC).

Solution

Let Y_{ij} denote the resposnse at the jth time point (t_{ij}) for the ith child. Define g_i=0 if girl, $g_i = 1$ if boy. We can write the model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + g_i(\eta_0 + \eta_1 t_{ij}) + e_{ij} = \beta_0 + g_i\eta_0 + t_{ij}(\beta_1 + \eta_1 g_i) + e_{ij}.$$

Load and pre-process the data as follows for easy handling.

```
library(nlme)

dental.data=read.table("../data/dental.txt",header=F)

# Rename columns
names(dental.data) <- c("obs","child","age","distance","gender")

# total number of individuals
m = max(dental.data$child)

# Create factor variables for use in gls()
dental.data$obs  = as.factor(dental.data$obs)
dental.data$child = as.factor(dental.data$child)
dental.data$gender = as.factor(dental.data$gender)  # 0 for girls and 1 for boys
head(dental.data)
```

```
##   obs child age distance gender
## 1   1     1   8     21.0      0
## 2   2     1  10     20.0      0
## 3   3     1  12     21.5      0
## 4   4     1  14     23.0      0
## 5   5     2   8     21.0      0
## 6   6     2  10     21.5      0
```

To investigate the covariance assumption, we consider the following covariance models.

```
# untructured
dental.un <- gls(distance ~ gender + age*gender, data=dental.data,
                 correlation=corSymm(form =  ~ 1 | child),
                 weights = varIdent(form = ~ 1 | age), method="ML")

# compound symmetry with equal variance for each time
dental.cs <- gls(distance ~ gender + age*gender, data=dental.data,
                 correlation = corCompSymm( , form= ~ 1 | child ),  method="ML")

# heterogeneous compound symmetry with UNequal variance for each time
dental.csh1 <- gls(distance ~ gender + age*gender, data=dental.data,
                 correlation = corCompSymm( , form= ~ 1 | child ),
                 weights = varIdent(form = ~ 1 | age), method="ML")

# autoregressive
dental.ar <- gls(distance ~ gender + age*gender, data=dental.data,
                 corr = corAR1(, form= ~ 1 | child),  method="ML")

# heterogeneous autoregressive
dental.arh1 <- gls(distance ~ gender + age*gender, data=dental.data,
                 corr = corAR1(, form = ~ 1 | child),
                 weight = varIdent(form = ~ 1 | age),  method="ML")

## aic and bic
aic <- AIC(dental.un, dental.cs, dental.csh1, dental.ar, dental.arh1)
bic <- BIC(dental.un, dental.cs, dental.csh1, dental.ar, dental.arh1)

# Display
cbind(aic, bic$BIC)
```

```
##             df      AIC  bic$BIC
## dental.un   14 447.4770 485.0269
## dental.cs    6 440.6391 456.7318
## dental.csh1  9 444.7166 468.8558
## dental.ar    6 452.6810 468.7738
## dental.arh1  9 456.7470 480.8862
```

Based on AIC/BIC, it seems the homogeneous compound symmetry (i.e., with equal variance for each time) fits the model better that our other choices. So we will go with this choice for the rest of the problem.

(ii) Given the choice of covariance model from (i) treat age as a categorical variable and fit a model which includes the effects of age, gender and their interactions. Determine whether the pattern of change over time is different for boys and girls.

Solution

We treat age as a caterorical variable. Define $I(\cdot)$ be indicator function. Now the model we consider is

$$Y_{ij} = \beta_0 + g_i\eta_0 + I(t_{ij} == 10)(\beta_1 + \eta_1 g_i) + I(t_{ij} == 12)(\beta_2 + \eta_2 g_i) + I(t_{ij} == 14)(\beta_3 + \eta_3 g_i) + e_{ij}.$$

```
## (ii) catergorical variable age
dental.data$fac.age = as.factor(dental.data$age)
model.fac.age = gls(distance ~ gender + fac.age*gender, data=dental.data,
                 correlation = corCompSymm(, form= ~ 1 | child ),  method="ML")
summary(model.fac.age)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: distance ~ gender + fac.age * gender
##   Data: dental.data
##        AIC      BIC    logLik
##   446.6329 473.4542 -213.3165
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | child
##  Parameter estimate(s):
##       Rho
## 0.6245473
##
## Coefficients:
##                        Value Std.Error   t-value p-value
## (Intercept)       21.181818 0.6915345 30.630167  0.0000
## gender1            1.693182 0.8983297  1.884811  0.0624
## fac.age10          1.045455 0.5992479  1.744611  0.0841
## fac.age12          1.909091 0.5992479  3.185812  0.0019
## fac.age14          2.909091 0.5992479  4.854570  0.0000
## gender1:fac.age10 -0.107955 0.7784458 -0.138680  0.8900
## gender1:fac.age12  0.934659 0.7784458  1.200673  0.2327
## gender1:fac.age14  1.684659 0.7784458  2.164131  0.0328
##
##  Correlation:
##                   (Intr) gendr1 fc.g10 fc.g12 fc.g14 g1:.10 g1:.12
## gender1           -0.770
## fac.age10         -0.433  0.334
## fac.age12         -0.433  0.334  0.500
## fac.age14         -0.433  0.334  0.500  0.500
## gender1:fac.age10  0.334 -0.433 -0.770 -0.385 -0.385
## gender1:fac.age12  0.334 -0.433 -0.385 -0.770 -0.385  0.500
## gender1:fac.age14  0.334 -0.433 -0.385 -0.385 -0.770  0.500  0.500
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.66200913 -0.63364313 -0.08238326  0.63847028  2.39297630
##
## Residual standard error: 2.20698
## Degrees of freedom: 108 total; 100 residual
```

P-value for the interaction between age and gender effect is 0.0626. The interaction is not significant at the level of significance 0.05, so the does not pattern of change over time is different for boys and girls. However, since the p-value is only 0.06, it seems that there might be some evidence that the patterns of change over time is different for boys and girls.

(iii) Use the estimated regression coefficients from (ii) to estimate the means in the two groups at ages 8 and 14.

Solution

We can obtain esetimated means based on estimated model coefficients:

```
model.fac.age$coefficients
```

```
##       (Intercept)           gender1         fac.age10         fac.age12
##        21.1818182         1.6931818         1.0454545         1.9090909
##         fac.age14 gender1:fac.age10 gender1:fac.age12 gender1:fac.age14
##         2.9090909        -0.1079545         0.9346591         1.6846591
```

Estimated means in Boys ($g_i = 1$) at ages 8 is: $\widehat{\beta}_0 + \widehat{\eta}_0$,

```
sum(model.fac.age$coefficients[c(1,2)])
```

```
## [1] 22.875
```

Estimated means in Boys ($g_i = 1$) at ages 14 is: $\widehat{\beta}_0 + \widehat{\eta}_0 + \widehat{\beta}_3 + \widehat{\eta}_3$,

```
sum(model.fac.age$coefficients[c(1,2, 5, 8)])
```

```
## [1] 27.46875
```

Estimated means in Girls ($g_i = 0$) at ages 8 is: $\widehat{\beta}_0$,

```
sum(model.fac.age$coefficients[c(1)])
```

```
## [1] 21.18182
```

Estimated means in Boys ($g_i = 0$) at ages 14 is: $\widehat{\beta}_0 + \widehat{\beta}_3$,

```
sum(model.fac.age$coefficients[c(1,5)])
```

```
## [1] 24.09091
```

(iv) Given the choice of model for the covariance from (i) treat age as a continuous variable and fit a model which includes the effect of a linear trend in age, gender, and their interaction. Plot the estimated mean for the two groups. Compare the results with those obtained (ii).

Solution

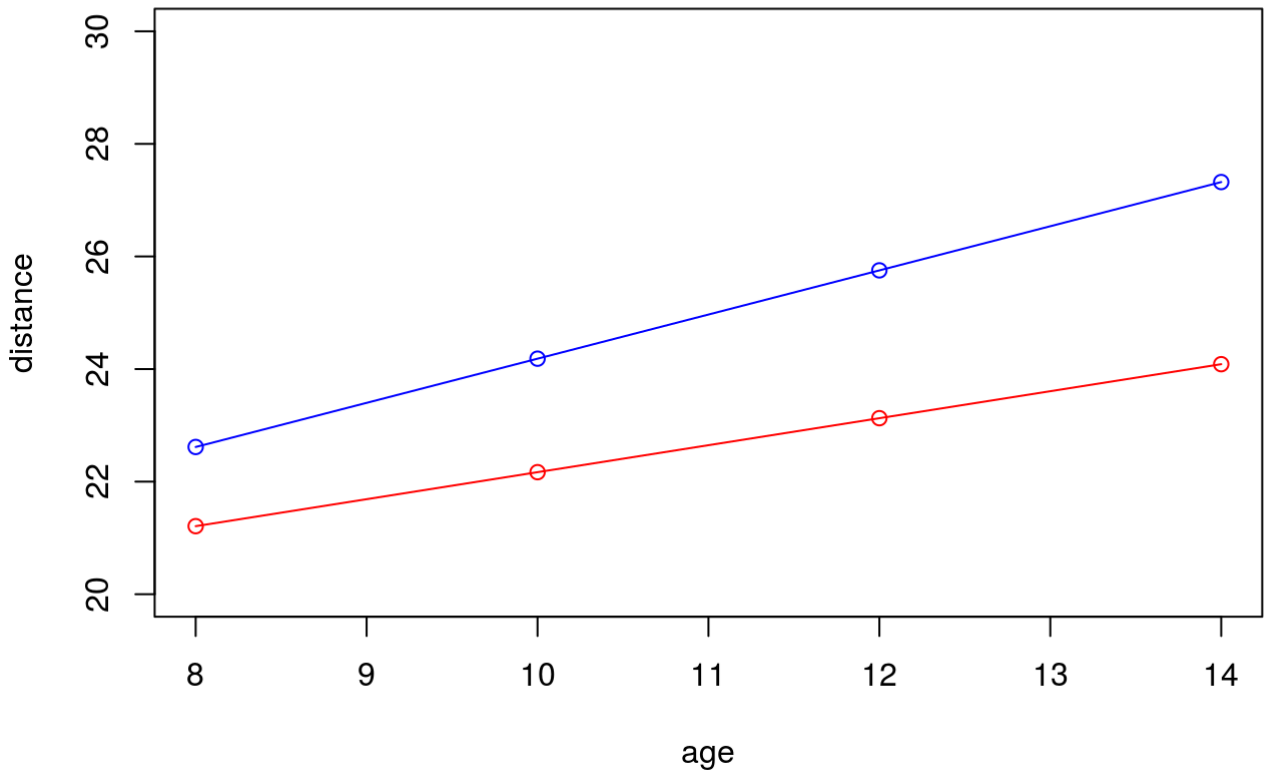
```
## (iv) continuouse variable age
full = gls(distance ~ gender + age*gender, data=dental.data,
           correlation = corCompSymm( , form= ~ 1 | child ),  method="ML")

reduced = gls(distance ~ gender + age, data=dental.data,
              correlation = corCompSymm( , form= ~ 1 | child ),  method="ML")

anova(full,reduced)
```

```
##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## full        1  6 440.6391 456.7318 -214.3195
## reduced     2  5 444.8565 458.2671 -217.4282 1 vs 2 6.217427  0.0126
```

```
## plot
girl.mean = full$coefficients[1] + full$coefficients[3]*c(8,10,12,14)
boy.mean = full$coefficients[1] + full$coefficients[2] + (full$coefficients[3] + full$co
efficients[4])*c(8,10,12,14)
plot(c(8,10,12,14), girl.mean, type="o", col="red", ylim=c(20,30), xlab="age", ylab="dis
tance")
lines(c(8,10,12,14), boy.mean, type="o", col="blue")
```



It seems that the patterns over time is different for boys and girls. We need to study if the interaction effect is significant. P-value for the interaction between age and gender effect is 0.0126 The interaction is significant at the level of significance 0.05. Thus, the patterns of change over time is different for boys and girls. We can investigate that models are different according to how to treat time covariate, age as categorical or continuous variable. The corresponding results could be different based on your model choice.

(v) Use the regression coefficients from (iv) and estimate the means in the two groups at ages 8 and 14.

Solution

We can obtain the estimated means based on estimated model coefficients as follows.

```
## Estimated coefficients
full$coefficients
```

```
## (Intercept)     gender1         age gender1:age
##  17.3727273  -1.0321023   0.4795455   0.3048295
```

```
## Estimated means in Boys at ages 8
boy8 = full$coefficients[1] + full$coefficients[2] + 8*(full$coefficients[3] + full$coef
ficients[4])
boy8
```

```
## (Intercept)
##    22.61562
```

```
## Estimated means in Boys at ages 14
boy14 = full$coefficients[1] + full$coefficients[2] + 14*(full$coefficients[3] + full$co
efficients[4])
boy14
```

```
## (Intercept)
##    27.32187
```

```
## Estimated means in Girls at ages 8
girl8 = full$coefficients[1] + 8*full$coefficients[3]
girl8
```

```
## (Intercept)
##    21.20909
```

```
## Estimated means in Girls at ages 14
girl14 = full$coefficients[1] + 14*full$coefficients[3]
girl14
```

```
## (Intercept)
##    24.08636
```

(vi) The 3rd measure (at age 12) on subject ID=20 is a potential outlier. Repeat the analyses in problems (i), (ii), and (iv) excluding the 3rd measure on subject ID=20. Do the substantive conclusions change?

Solution:

```
## remove a potential outlier
time = rep(c(1,2,3,4), m)
dental.data$time = time
dental.data2 = dental.data[-79, ]


## Repeat (i)
# un
dental.un <- gls(distance ~ gender + age*gender, data=dental.data2,
                 correlation=corSymm(form =  ~ time | child),
                 weights = varIdent(form = ~ time | age), method="ML")
# cs
dental.cs <- gls(distance ~ gender + age*gender, data=dental.data2,
                 correlation = corCompSymm(, form= ~ time | child ),  method="ML")
# heterog cs
dental.csh1 <- gls(distance ~ gender + age*gender, data=dental.data2,
                 correlation = corCompSymm( , form= ~ time | child ),
                 weights = varIdent(form = ~ time | age), method="ML")
# autoregressive
dental.ar <- gls(distance ~ gender + age*gender, data=dental.data2,
                 corr = corAR1(, form= ~ time | child),  method="ML")
# hetero autoregressive
dental.arh1 <- gls(distance ~ gender + age*gender, data=dental.data2,
                 corr = corAR1(, form = ~ time | child),
                 weight = varIdent(form = ~ time | age),  method="ML")

## aic and bic
aic=AIC(dental.un, dental.cs, dental.csh1, dental.ar, dental.arh1)
bic=BIC(dental.un, dental.cs, dental.csh1, dental.ar, dental.arh1)
cbind(aic, bic$BIC)
```

```
##             df      AIC  bic$BIC
## dental.un   14 413.2993 450.7189
## dental.cs    6 417.7575 433.7944
## dental.csh1  9 421.8926 445.9481
## dental.ar    6 417.8228 433.8598
## dental.arh1  9 420.6116 444.6670
```

In AIC, common unstructured is the best assumption and common compound symmetry or AR(1) is the best choice in BIC. First, we use common compound symmetry and repeat the analysis in problem (ii) and (iv)

```
## repeat (ii)
model.fac.age = gls(distance ~ gender + fac.age*gender, data=dental.data2,
                 correlation = corCompSymm(, form= ~ time | child ),  method="ML")
#summary(model.fac.age)

## test
reduced.model = gls(distance ~ gender + fac.age, data=dental.data2,
           correlation = corCompSymm(, form= ~ time | child ),  method="ML")
anova(model.fac.age, reduced.model)
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model.fac.age     1 10 421.3762 448.1045 -200.6881
## reduced.model     2  7 424.3705 443.0803 -205.1853 1 vs 2 8.994266  0.0294
```

```
## repeat (iv)
full = gls(distance ~ gender + age*gender, data=dental.data2,
           correlation = corCompSymm( , form= ~ 1 | child ),  method="ML")
reduced = gls(distance ~ gender + age, data=dental.data2,
              correlation = corCompSymm( , form= ~ 1 | child ),  method="ML")
anova(full,reduced)
```

```
##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## full        1  6 417.7575 433.7944 -202.8787
## reduced     2  5 422.6329 435.9971 -206.3165 1 vs 2 6.875454  0.0087
```

Second, we use AR(1) and repeat the analysis in problem (ii) and (iv)

```
## repeat (ii)
model.fac.age = gls(distance ~ gender + fac.age*gender, data=dental.data2,
                 corr = corAR1(, form= ~ time | child),  method="ML")
#summary(model.fac.age)

## test
reduced.model = gls(distance ~ gender + fac.age, data=dental.data2,
           corr = corAR1(, form= ~ time | child),  method="ML")
anova(model.fac.age, reduced.model)
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model.fac.age     1 10 420.2570 446.9853 -200.1285
## reduced.model     2  7 420.0708 438.7806 -203.0354 1 vs 2 5.813811   0.121
```

```
## repeat (iv)
full = gls(distance ~ gender + age*gender, data=dental.data2,
           corr = corAR1(, form= ~ time | child),  method="ML")
reduced = gls(distance ~ gender + age, data=dental.data2,
              corr = corAR1(, form= ~ time | child),  method="ML")
anova(full,reduced)
```

```
##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## full        1  6 417.8228 433.8598 -202.9114
## reduced     2  5 419.1016 432.4657 -204.5508 1 vs 2 3.278739  0.0702
```

When we use the compound symmetry model, the results are similar as the previous ones. However, when we use AR(1) as the covariance model, interaction effect is not significant, which indicates that the results are influenced by the outlier and the choice of covariance model.

(vii) Given the results of all the previous analyses, what conclusions can be drawn about the gender differences in patterns of dental growth.

The interaction between age and gender is significant before and after removing the potential outlier when we use compound symmetry covariance assumption. Thus, we can make a conclusion that the patterns of dental growth are different for boys and girls.

Problem 2 (35 points)

Exposure to lead can produce a variety of adverse health effects in infants and children, including hyperactivity, hearing or memory loss, learning disabilities, and damage to the nervous system. Although the use of lead as a gasoline additive has been discontinued in the US, so that airborne lead levels have been reduced dramatically, a small percentage of children continue to be exposed to lead at levels that can produce such health problems. Much of this exposure is due to deteriorating lead-based paint that may be chipping and peeling in older homes. Lead-based paint in housing was banned in the US in 1978; however, many older homes (built pre-1978) do contain lead-based paint, and chips and dust can be ingested by young children living in these homes during normal teething and hand-to-mouth behavior. This is especially a problem among children in deteriorating, inner-city housing. The US Centers for Disease Control and Prevention (CDC) has determined that children with blood levels above 10 micrograms/deciliter (μg/dL) of whole blood are at risk of adverse health effects.

Luckily, there are so-called chelation treatments that can help a child to excrete the lead that has been ingested. The researchers were interested in evaluating the effectiveness of one such chelating treatment, succimer, in children who had been exposed to what the CDC views as dangerous levels of lead. They conducted the following study. 120 children aged 12{36 months with confirmed blood lead levels of $> 15\mu$g/dL and ; 40μg/dL in a large, inner-city housing project were identified; these lead levels are above the at-risk threshold determined by the CDC.

A clinic was set up in the housing project staffed by personnel from the city's Department of Public Health. The personnel randomized the children into three groups: 40 children were assigned at random to receive a placebo (an inactive agent with no lead-lowering properties), 40 children were assigned at random to receive a low dose of succimer, and 40 children were assigned at random to receive a higher dose of succimer. Blood lead levels were measured at the clinic for each child at baseline (time 0), prior to initiation of the assigned treatments. Then, assigned treatment was started, and, ideally, each child was to return to the clinic at weeks 2, 4, 6, and 8. At each visit, blood lead level was measured for each child.

The data are available in the file [lead.dat.txt] (../data/lead.full.txt). The data are presented in the form of one data record per observation; the columns of the data set are as follows:

1 Child id

2 Indicator of age (= 0 if ≤ 24 months; = 1 if > 24 months)

3 Gender indicator (= 0 if female, = 1 if male)

4 Week

5 Blood lead level (μg/dL)

6 Treatment indicator (= 1 if placebo, = 2 if low dose, = 3 if higher dose)

```
lead <- read.table("lead.full.txt", header = F)
colnames(lead) = c("id", "ind.age", "sex", "week", "blood", "trt")
head(lead)
```

```
##   id ind.age sex week blood trt
## 1  1       0   1    0  31.8   1
## 2  1       0   1    2  31.6   1
## 3  1       0   1    4  39.9   1
## 4  1       0   1    6  40.5   1
## 5  1       0   1    8  48.3   1
## 6  2       0   0    0  24.5   1
```

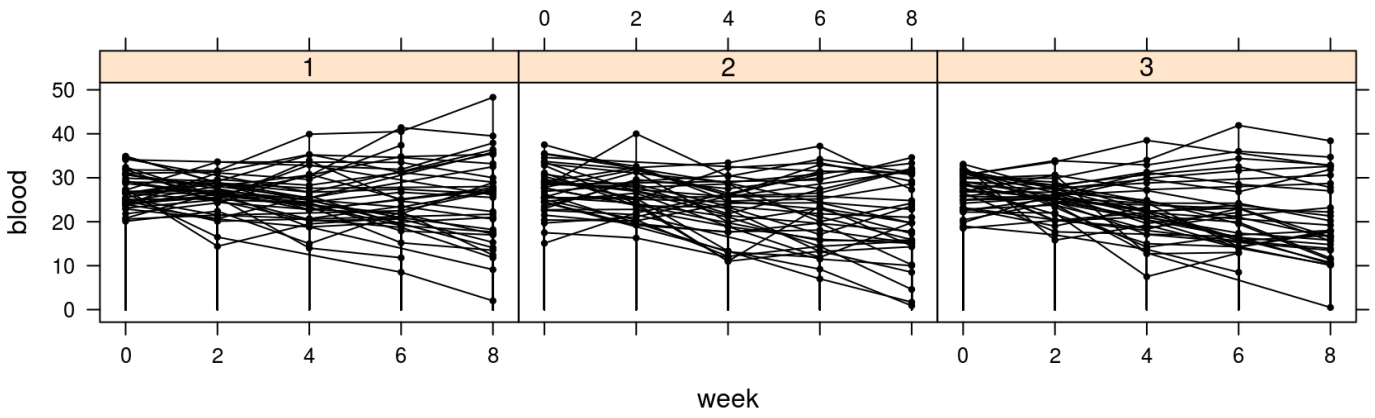
The investigators had several questions of interest. Broadly stated, the primary focus was on whether succimer, in either low- or high-dose form is effective over an eight week period in reducing blood lead levels in this population of children. They were also interested in whether blood lead levels in this population are associated with the age and/or gender of the child, and whether the effectiveness of succimer in reducing blood lead levels is associated with either or both of these factors.

Now solve the following problems.

(a) Draw profile plots for each of the three treatment groups (keep the limits of axes the same for all the plots for a fair comparison). Comment on any pattern you see in these plots.

[Hint: You will notice that, although all children were observed at baseline, some children are missing some of the intended subsequent lead level measurements. This might be because some children were unable to come to the clinic for an assigned visit because their caregiver was unable to bring them. The investigators interviewed these children's caregivers and felt comfortable assuming that the the inability of some children to show up for some visits was not related to which treatment they were taking or how they were doing on their assigned treatment. As we will discuss later in the course, the validity of an analysis may be compromised if missingness is related to the thing under study in certain ways.]

```
library(lattice)
lead$trtgroup <- factor(lead$trt)
xyplot(blood ~ week | trtgroup, data = lead, groups = id, type = c("b", "h"), col="blac
k", pch=19, cex=0.4)
```



(b) Analyze the above data to best address the questions that the investigators are interested in. Specifically, for each group fit a model that assumes a linear trend in time (week), age (i.e. age indicator), gender and their interaction. In your analysis consider the following model for the covariance structure:

i. Independence in both groups with the same variance

ii. Homogeneous compound symmetry, same in all groups and then different for each group

iii. Unstructured, same in all groups and then different for each group.

```
## Defining factors for easy analysis
## factor
child = factor(lead$id)
gender = factor(lead$sex, labels = c("female", "male"))
age = factor(lead$ind.age, labels = c("lower24", "upper24"))
group = factor(lead$trt, labels=c("placebo", "low", "higher"))

## time variable to account for the missingness
time = lead$week/2 + 1
lead = cbind(lead, child, gender, age, group, time)
head(lead)
```

```
##   id ind.age sex week blood trt trtgroup child gender     age   group time
## 1  1       0   1    0  31.8   1        1     1   male lower24 placebo    1
## 2  1       0   1    2  31.6   1        1     1   male lower24 placebo    2
## 3  1       0   1    4  39.9   1        1     1   male lower24 placebo    3
## 4  1       0   1    6  40.5   1        1     1   male lower24 placebo    4
## 5  1       0   1    8  48.3   1        1     1   male lower24 placebo    5
## 6  2       0   0    0  24.5   1        1     2 female lower24 placebo    1
```

```
## mean model
model.form <- blood ~ group + age + gender + group:age + group:gender + age:gender +  gr
oup:age:gender + week +
                + group:week + group:age:week + group:gender:week+group:age:gender:week

## Independence in both grous with the same variance
lead.ind <- gls(model.form, data=lead,  method="ML")


# Homogeneous Compound Symmetry
lead.cs <- gls(model.form, data=lead,
               correlation=corCompSymm(form =  ~ time | child ), method="ML")

## Compound Symmetry, different variance for each group
lead.cs2 <- gls(model.form, data=lead,
               correlation=corCompSymm(form =  ~ time | child ),
               weights = varIdent(form = ~ time | group),  method="ML")

## Unstructured, common variances for each group
lead.un <- gls(model.form, data=lead,
               correlation=corSymm(form =  ~ time | child),
               weights = varIdent(form = ~ time | week ) , method="ML")

## Unstructured, different variances for each group
lead.un2 <- gls(model.form, data=lead,
               correlation=corSymm(form =  ~ time | child),
               weights = varIdent(form = ~ time | week*group ) , method="ML")

## AIC and BIC (Smallest is the best)
aic = AIC(lead.ind, lead.cs, lead.cs2, lead.un, lead.un2)
bic = BIC(lead.ind, lead.cs, lead.cs2, lead.un, lead.un2)
cbind(aic, bic$BIC)
```

```
##          df      AIC  bic$BIC
## lead.ind 25 3429.610 3536.243
## lead.cs  26 3310.438 3421.336
## lead.cs2 28 3314.373 3433.802
## lead.un  39 3086.238 3252.585
## lead.un2 49 3098.291 3307.291
```

Make a table of AIC and BIC values for these models. Based on these results, select the model for which you think the evidence in the data is strongest, explaining your answer. Based on your

selected model from the above analysis, clearly write out the mathematical model for this data.

According to AIC and BIC criteria, we use the common unstructured model for the rest of the problem.

(c) Based on your chosen covariance model,

i. Study if gender has a significant effect.

ii. Use the model resulted in (i) and study if age has significant effect.

iii. Use the model resulted in (ii) and study whether the rate of change of the lead level is different across groups.

To test this hypothesis, we will use likelihood ratio test.

Test overall gender effect

```
# full model (with all terms involving gender present in the model)
## mean model
model.form <- blood ~ group + age + gender + group:age + group:gender + age:gender +  gr
oup:age:gender + week +
                  + group:week + group:age:week + group:gender:week+group:age:gender:week
full.model <- gls(model.form, data=lead,
               correlation=corSymm(form =  ~ time | child),
               weights = varIdent(form = ~ time | week ) , method="ML")


# reduced model (with all terms involving gender ABSENT in the model)
model.nogender <- blood ~ group + age + group:age + week + group:week + group:age:week
reduced.model = gls(model.nogender, data=lead,
               correlation=corSymm(form =  ~ time | child),
               weights = varIdent(form = ~ time | week ) , method="ML")

# LRT
anova(full.model, reduced.model)
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## full.model        1 39 3086.238 3252.585 -1504.119
## reduced.model     2 27 3069.236 3184.399 -1507.618 1 vs 2 6.997767  0.8578
```

From the output, the LRT test statistic is $T = 6.99$ with p-value 0.86. There does not seem to be evidence at any reasonable level of significance (e.g., 0.05) in the data to claim that there are gender effects in either intercept or slope.

Test age effect

```
# full model
full.model <- gls(model.nogender, data=lead,
               correlation=corSymm(form =  ~ time | child),
               weights = varIdent(form = ~ time | week ) , method="ML")

# reduced model (no age effect)
model.noage <- blood ~ group + week + group:week
reduced.model = gls(model.noage, data = lead,
               correlation=corSymm(form =  ~ time | child),
               weights = varIdent(form = ~ time | week ) , method="ML")

anova(full.model, reduced.model)
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## full.model        1 27 3069.236 3184.399 -1507.618
## reduced.model     2 21 3124.627 3214.198 -1541.313 1 vs 2 67.39125  <.0001
```

From the output, the LRT test statistic is $T = 67.4$ with p-value near zero. There is clearly strong evidence suggesting that mean lead level is associated with age.

Test rate of change of the lead level is different across groups

```
# full model
model.nogender <- blood ~ group + age + group:age + week + group:week + group:age:week
full.model <- gls(model.nogender, data=lead,
               correlation=corSymm(form =  ~ time | child),
               weights = varIdent(form = ~ time | week ) , method="ML")

# reduced model (same slope of week in all the groups)
model.slope <- blood ~ group + age + group:age + week
reduced.model = gls(model.slope, data=lead,
               correlation=corSymm(form =  ~ time | child),
               weights = varIdent(form = ~ time | week ) , method="ML")

anova(full.model, reduced.model)
```

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## full.model        1 27 3069.236 3184.399 -1507.618
## reduced.model     2 22 3068.789 3162.625 -1512.394 1 vs 2 9.553018  0.0889
```

From the output, the test statistic is $T = 9.55$ with a p-value 0.08. At level of significance 0.05, there is not enough evidence to suggest that the slopes differs from the others.

(d) Give a brief summary of your findings for this study.

Overall, we find that gender does not semm to have any effect on lead level, but age and treatment do.

Problem 3 (20 points)

Consider the hip replacement study (see Example C in the lecture [Models for mean and covariance] (../Lecture08_LDA_Modeling_and_Estimation.html)). We discussed that a quadratic model in `time` would be appropriate here for each group along with a linear effect of `age`. Answer the follwoing questions (no need for data analysis).

(a) Let t_{ij} denote the j-th observation for the i-th individual. Write down a mathematical model for this data *assuming that `age` has the same effect for both the groups*.

[Hint: there should be 7 regression coefficients]

Model:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + G_i(\beta_3 + \beta_4 t_{ij} + \beta_5 t_{ij}^2) + \beta_6 age_{ij} + e_{ij}.$$

(b) Let $\boldsymbol{Y}_i$ be the data vector of the i-th individual. Express your model in (a) in terms of a desigm matrix $\boldsymbol{X}_i$ and a parameter $\boldsymbol{\beta}$. Give the form of $\boldsymbol{X}_i$ for individuals in each group.

[Hint: $\boldsymbol{\beta}$ should be of length 7 and each $\boldsymbol{X}_i$ should have 7 columns]

Group 1 with $G = 0$

$$\begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{bmatrix} = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 \\ \vdots & & \\ 1 & t_{i1} & t_{i1}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_{i1} \\ \vdots \\ e_{im_i} \end{bmatrix}$$

Group 1 with $G = 1$

$$\begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{bmatrix} = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 \\ \vdots & & \\ 1 & t_{i1} & t_{i1}^2 \end{bmatrix} \begin{bmatrix} \beta_0 + \beta_3 \\ \beta_1 + \beta_4 \\ \beta_2 + \beta_5 \end{bmatrix} + \begin{bmatrix} e_{i1} \\ \vdots \\ e_{im_i} \end{bmatrix}$$

(c) We might ask the question whether the study was carried out properly in the sense that the individuals were similar on average at baseline. In the context of your model in (a) and (b), write down the null and alternative hypothesis that addresses this question. Express your hypotheses in terms of $\boldsymbol{L\beta}$ for some appropriate matrix $\boldsymbol{L}$, giving the form of $\boldsymbol{L}$.

[Hint: This $\boldsymbol{L}$ will have one row and 7 columns.]

Here we are interested in $\beta_3 = 0$ or not. So

$$L = [0, 0, 0, 1, 0, 0, 0]$$

(d) The next question was whether all groups tend to have mean hematocrit profiles that change at constant rates (i.e., only linear trends, but not quadratic, that are possibly different in each

group) or whether at least one of the groups exhibits an "acceleration" (i.e., are one or more quadratic terms nonzero). In the context of your model in (a) and (b), write down the null and alternative hypothesis that addresses this question. Express your hypotheses in terms of $L\beta$ for some appropriate matrix L, giving the form of L.

[Hint: This L will have two rows and 7 columns.]

Here we test $\beta_2 = 0$ and $\beta_5 = 0$

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

(e) Finally, we ask the question whether the mean hematocrit profiles show an identical pattern of change across time for all groups. Assuming your model in (a) and (b), write down the null and alternative hypothesis that addresses this issue. Express your hypotheses in terms of $L\beta$ for some appropriate matrix L, giving the form of L.

[Hint: This L will have two rows and 7 columns. Note we are interested in pattern of change, not the baseline itself].

Here we test $\beta_4 = 0$ and $\beta_5 = 0$

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

Problem 4 (10 points)

Consider data that are balanced, so that each experimental unit is observed at the same m times $t_1, \ldots, t_m$.

(a) Suppose that $m = 4$ and that the times are one unit apart. Write down the correlation matrix for a single experimental unit Y_i when the covariance structure is that of an autoregressive model of order 1, AR(1), with $\rho = 0.6$.

```
##       [,1] [,2] [,3]  [,4]
## [1,] 1.000 0.60 0.36 0.216
## [2,] 0.600 1.00 0.60 0.360
## [3,] 0.360 0.60 1.00 0.600
## [4,] 0.216 0.36 0.60 1.000
```

(b) For the same situation as (a), suppose that Y_i has a missing value at t_3. Write down the 3×3 correlation matrix of Y_i.

```
##       [,1] [,2]  [,3]
## [1,] 1.000 0.60 0.216
## [2,] 0.600 1.00 0.360
## [3,] 0.216 0.36 1.000
```