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).

(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.

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

(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).

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

(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?

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

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 $> 15\mu\mathrm{g/dL}$ and ; $40\mu\mathrm{g/dL}$ 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.]

(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.

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.

(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.

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

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]

(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]

(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.]

(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 $\boldsymbol{L\beta}$ for some appropriate matrix $\boldsymbol{L}$, giving the form of $\boldsymbol{L}$.

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

(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 $\boldsymbol{L\beta}$ for some appropriate matrix $\boldsymbol{L}$, giving the form of $\boldsymbol{L}$.

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

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 $\boldsymbol{Y}_i$ when the covariance structure is that of an autoregressive model of order 1, AR(1), with $\rho = 0.6$.

(b) For the same situation as (a), suppose that $\boldsymbol{Y}_i$ has a missing value at t_3. Write down the 3×3 correlation matrix of $\boldsymbol{Y}_i$.