ST 437/537 Homework 4, Spring 2018

ST 537: Q1 and Q2. ST 437: Q1 parts (a), (c), (d) & Q2 i. and ii.

1. Orthodontist data. Let Y_{ij} denote the jth distance (mm) from the center of the pituitary to the pterygomaxillary fissure on the ith child at time (age) t_{ij}. Define $g_i = 0$ if girl, $g_i = 1$ if boy. Consider the following model:

$$Y_{ij} = \beta_0 + \beta_{0g} g_i + (\beta_1 + \beta_{1g} g_i) t_{ij} + \epsilon_{ij},$$

where $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{i4})^T$ is assumed to be IID $N(0_4, \Sigma_i(\theta))$, and $\Sigma(\theta)$ is parametric covariance model.

   - Do an exploratory analysis of the data. Are there outliers ? Report your findings.
   - Study which covariance model is most appropriate for this data. For this part it's sufficient to examine the following covariance structures: compound symmetric (w/out heterogeneous version), autoregressive (w/out heterogeneous version), unstructured.

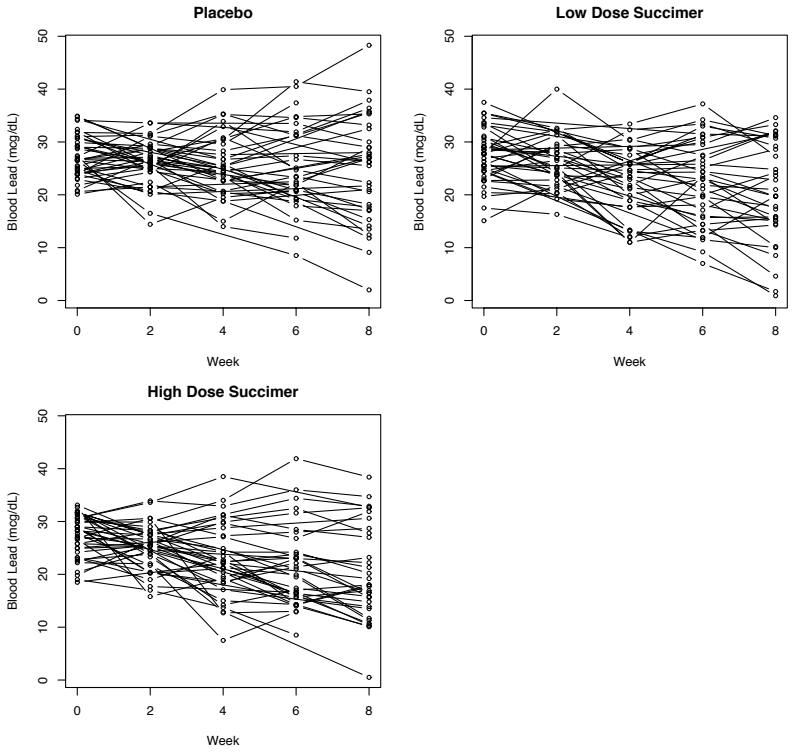
     Irrespective of your answer to the previous part, assume the residual has covariance described by compound symmetric with constant variance.
   - We want to study whether there is significant interaction between Age and Gender. Investigate this problem by first assuming that Age t_{ij} is a categorical variable (make sure you specify this in `R`) .
   - We want to study whether there is significant interaction between Age and Gender. Investigate this problem by treating Age t_{ij} is a continuous variable (make sure you specify this in `R`) . Which model do you prefer? Justify your reasoning.
   - Are your conclusions affected by the presence of outliers ? Carry out the necessary additional investigation to address this problem.

2. 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 μg/dL and , 40 mug/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.

Figure 1: *Blood lead levels for three groups of children*



The data are available in the file `lead.dat.txt` on the class web page (Lead data). 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 $\leq$ 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)

You will notice in Figure 2 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.

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. They postulated the following model.

Let Y_{ij} denote the jth lead level measurement on the ith child at time t_{ij} for that child, $j = 1, \ldots, n_i$. Note that the t_{ij} for each child and n_i may be different. Define $a_i = 0$ if subject i's age is $\leq$ 24 months and $a_i = 1$ if age is > 24. Let g_i indicate the gender of child i ($g_i = 0$ if female, =1 if male). The model they considered is

$$\begin{aligned}
Y_{ij} &= (\beta_0 + \beta_{0a}a_i + \beta_{0g}g_i + \beta_{0ag}a_ig_i) + (\beta_1 + \beta_{1a}a_i + \beta_{1g}g_i + \beta_{1ag}a_ig_i)t_{ij} + \epsilon_{ij} \text{ placebo} \\
Y_{ij} &= (\beta_0 + \beta_{0a}a_i + \beta_{0g}g_i + \beta_{0ag}a_ig_i) + (\beta_2 + \beta_{2a}a_i + \beta_{2g}g_i + \beta_{2ag}a_ig_i)t_{ij} + \epsilon_{ij} \text{ low dose} \\
Y_{ij} &= (\beta_0 + \beta_{0a}a_i + \beta_{0g}g_i + \beta_{0ag}a_ig_i) + (\beta_3 + \beta_{3a}a_i + \beta_{3g}g_i + \beta_{3ag}a_ig_i)t_{ij} + \epsilon_{ij} \text{ high dose}
\end{aligned} \tag{1}$$

Thus, this is a rather complicated model!

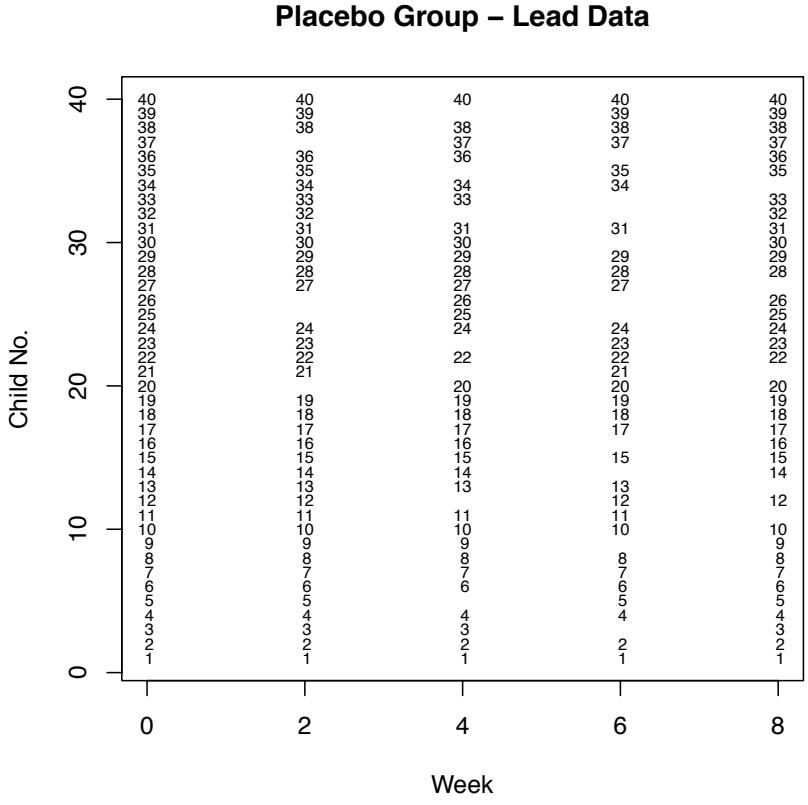
i. Which covariance model is appropriate to use in order to describe the covariance in the data

Here are the different covariance structures to consider:

(i) Independence in both groups with the same variance (this may be accomplished by not including a `repeated` statement at all)
(ii) Homogeneous compound symmetry, same in all groups and then different for each group
(iii) Homogeneous AR(1), same in all groups and then different for each group
(iv) Unstructured (`type=un`), 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. *Adopt the covariance model you think is best for the rest of the problem.*

Figure 2: *Shows missing visits for the placebo group.*



ii. Study if gender has a significant effect

iii. Study if age has a significant effect

iv. Study whether the rate of change of the lead level is different across groups.