1. Consider data that are balanced, so that each experimental unit is observed at the same \( n \) times \( t_1, \cdots, t_n \).

(a) Suppose that \( n = 4 \) 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 \( \rho = 0.6 \).

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

(c) Now suppose that \( n = 4 \) but the times are at 0.5, 1.0, 2.0, and 5.0. For a Markov covariance structure with \( \rho = 0.6 \) (see p. 224), write down the correlation matrix of \( Y_i \).

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 (\( \mu g/dL \)) 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 > 15µg/dL and, 40 µg/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 1, 2, 4 and 8. At each visit, blood lead level was measured for each child.

The data are available in the file lead.dat on the class web page. 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 (= 0 if placebo, = 1 if low dose, = 2 if higher dose)

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 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 $j$th lead level measurement on the $i$th child at time $t_{ij}$ for that child, $j = 1, \cdots, n_i$. 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 (1) they considered is:

For placebo group:

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

For low dose group:

$$Y_{ij} = (\beta_0 + \beta_{0a}a_i + \beta_{0g}g_i + \beta_{0ag}a_ig_i) + (\beta_2 + \beta_{2a}a_i + \beta_{2g}g_i + \beta_{2ag}a_ig_i)t_{ij} + \epsilon_{ij},$$

For high dose group:

$$Y_{ij} = (\beta_0 + \beta_{0a}a_i + \beta_{0g}g_i + \beta_{0ag}a_ig_i) + (\beta_3 + \beta_{3a}a_i + \beta_{3g}g_i + \beta_{3ag}a_ig_i)t_{ij} + \epsilon_{ij}.$$

Note the following features of model (1):

- For each group, there is a common intercept term

$$\beta_0 + \beta_{0a}a_i + \beta_{0g}g_i + \beta_{0ag}a_ig_i$$

such that mean lead level equals this at week 0. Thus, the model assumes that mean lead level at baseline (before treatment) in this population of children may be associated with age and gender of the child in a way that is different depending on the age-gender combination. Moreover, because the term is identical in all groups, it assumes that these features of baseline mean lead level are same in all three groups.
Similarly, for each group k, there is a slope term multiplying week given by

\[ \beta_k + \beta_{ka}a_i + \beta_{kg}g_i + \beta_{kag}a_i g_i. \]

Thus, the model assumes that the way in which mean lead level changes after baseline (i) possibly depends on the age and gender of the child in a way that is different depending on the age-gender combination; and (ii) the slope and the way it depends on age and gender are possibly different for each group.

You will want to run one SAS program repeatedly, adding on to the program and rerunning it as you work through the following problems. Turn in the final program that does everything required and its output. Use \texttt{PROC MIXED} to carry out all model fits. \textit{Whenever you carry out a hypothesis test, report the p-value, not just whether it rejects or not.}

(a) Make the spaghetti plots for the blood lead levels for each of the 3 groups, separately. Why do you think that the investigators assumed that the intercept is the same in all three groups? Explain.

(b) The first step in the analysis is to investigate the nature of variation, as the validity of subsequent inferences depends on an having an appropriate model for covariance structure. Fit model (1) for the following covariance structures. \textit{Hint:} It will take some figuring to determine an appropriate \texttt{model} statement to fit (1). The considerations in Section 8.9 apply, but are harder. You can make things easier by noting that, because age and gender are binary variables coded as 0-1, you can just leave them as numeric variables (do not include them in the \texttt{class} statement). The treatment indicator can be declared to be a \texttt{class} variable. Try applying the rules in Section 8.9 under these conditions to arrive at your \texttt{model} statement. Here are the different covariance structures to fit.

1. Independence in both groups with the same variance (this may be accomplished by not including a repeated statement at all)
2. Homogeneous compound symmetry, same in both groups and then different in both groups
(iii) Homogeneous one-dependent, same in both groups and then different in both groups

(iv) Homogeneous AR(1), same in both groups and then different in both groups

(v) Unstructured, same in both groups and then different in both groups.

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.*

For the unstructured covariance, have your program print out the estimated covariance and correlation matrix for child 1 (complete data) and child 3 (missing data) in group 1 (placebo) to verify you have handled the missing values correctly. Print out the estimated covariance and correlation matrix for child 44 in group 2 (low dose) and child 82 in group 3 (high dose) as well. Does the selected covariance structure make sense when looking at the unstructured estimates?

(c) Previous research on lead exposure says that while age of a child may be implicated in lead levels (older children have had more time to be exposed), gender of the child is not associated with lead levels in this population of children, nor is it associated with the progression of lead levels over time, with or without treatment. Based on these results, the investigators were hoping to simplify model (1) to have no dependence on gender in either the intercept or slope terms. Write down an appropriate null hypothesis that formalizes this in terms of the model in (1), and express it in the form \( H_0 : L\beta = 0 \), defining \( \beta \) and \( L \). Test your hypotheses two ways:

- Re-fit the model in (1) using your favored choice of covariance structure from (c) and include a **contrast** statement corresponding to the matrix \( L \). Write down the value of the test statistic \( T_L \) and the associated p-value from the output.

- Fit the reduced model implied by your null hypothesis and, based on its output and that from the full model (1), obtain (by hand) the likelihood ratio test statistic \( T_{LRT} \). (Recall this requires using **method=ml**.) State the degrees of freedom for this test statistic and the appropriate \( \chi^2 \) critical value from your favorite \( \chi^2 \) table or the p-value from your
favorite software package.

In each case, state the conclusion you draw from the test. Is there sufficient evidence in these data to refute the investigators claim?

(d) Write down the reduced model implied by $H_0$ in (c). Based on a fit of this model, do whatever you think necessary to address in the context of this model the question of whether there is an association of mean lead level with age in this population of children (i.e., at baseline). (That is, state appropriate null and alternative hypotheses and provide numerical evidence in the form of a test statistic and p-value, interpreting these.) Give estimates (and associated standard errors) for the mean lead level for younger ($\leq 24$ months old) and for older ($> 24$ months) children.

(e) Is there evidence to suggest that the age of a child is associated with the change in mean lead level for any of the treatments? Based on a fit of the model in (e), do whatever you think necessary to address this question. (That is, state appropriate null and alternative hypotheses and provide numerical evidence in the form of a test statistic and p-value, interpreting these.)

(f) Based on the fit of whatever model you feel is most relevant, address the general question: Is there a difference in rate of change of mean lead level over time among the three groups? (That is, state appropriate null and alternative hypotheses and provide numerical evidence in the form of a test statistic and p-value, interpreting these.)