.

ST 732, HOMEWORK 5, SPRING 2007

1. Recall the lead exposure study from Homework 3, Problem 3. Consider model (1) in the statement of that problem, which is repeated here for convenience: 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 ≤ 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 is

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

where the ϵ_{ij} have mean zero. Thus, in this model, if child i is in treatment group k, mean lead level at time t_{ij} is modeled directly as

$$E(Y_{ij}) = (\beta_0 + \beta_{0a}a_i + \beta_{0g}g_i + \beta_{0ag}a_ig_i) + (\beta_k + \beta_{ka}a_i + \beta_{kg}g_i + \beta_{kag}a_ig_i)t_{ij}. \tag{2}$$

If we adopt the usual notation, then, we can write this model in the form

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})'$,

$$\boldsymbol{\beta} = (\beta_0, \beta_{0a}, \beta_{0g}, \beta_{0ag}, \beta_1, \beta_{1a}, \beta_{1g}, \beta_{1ag}, \beta_2, \beta_{2a}, \beta_{2g}, \beta_{2ag}, \beta_3, \beta_{3a}, \beta_{3g}, \beta_{3ag})',$$

and the rows of $\boldsymbol{X}_i\boldsymbol{\beta}$ have elements for $j = 1, \ldots, n_i$ as in (2). When we specify models like this, we postulate a model for $\text{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_i$ *directly*. This model is an example of a "population-averaged" model (see p. 331 of the notes).

As did the investigators in Problem 3 of Homework 4, we can instead take a "subject-specific" approach to modeling these data and assume that each child has his/her own "inherent" straight-line trajectory, i.e.,

$$Y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + e_{ij}, \tag{3}$$

where e_{ij} has mean zero (the individual first stage model). At the population stage, we may assume that the child-specific intercepts β_{0i} and slopes β_{1i} for a child who was in treatment group k satisfy

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

$$\beta_{1i} = \beta_k + \beta_{ka}a_i + \beta_{kg}g_i + \beta_{kag}a_ig_i + b_{1i},$$

where the random effect vector $\boldsymbol{b}_i = (b_{0i}, b_{1i})'$ has mean $\boldsymbol{0}$.

(a) By substituting the above expressions for β_{0i} and β_{1i} into (3) and collecting terms, give the expression for Y_{ij} that is implied by the subject-specific model for a child assigned to treatment k. Based on this expression, demonstrate that the subject-specific model implies the assumption that $E(Y_{ij})$ is equal to the expression in (2).

(b) With $\boldsymbol{\beta}$ defined as above, the expressions for β_{0i} and β_{1i} above may be written compactly in the form $\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i$, where $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i})'$. Write down the form of $\boldsymbol{A}_i$ for each treatment

group. Then rewrite (3) and this model for $\boldsymbol{\beta}_i$ together in the form $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$, where $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})'$, defining $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ for all possible i (i.e., children in all three treatment groups). Note that, because of the result in (a), this $\boldsymbol{X}_i$ and the $\boldsymbol{X}_i$ for model (1) are *identical*.

(c) Thus, we have the result discussed on p. 331 of the notes: the main difference between population-averaged model (1), which can be written as $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, and the subject-specific model, written as $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$ is that the latter imposes a *specific structure* on the "deviations" from $\boldsymbol{X}_i\boldsymbol{\beta}$. In fact, we could write the latter as $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ with the stipulation that $\boldsymbol{\epsilon}_i$ has the specific form $\boldsymbol{\epsilon}_i = \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$. As we know, it is this specific structure that imposes the form of $\text{var}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i$ for linear mixed effects models.

What this shows is that we may view model (1) and the subject-specific models as *competing models* in the sense that they assume the *same* form of the mean of a data vector but *differ* in terms of how the covariance matrix $\text{var}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i$ is represented. From this perspective, if our primary interest is in inference on $\boldsymbol{\beta}$ in a given mean model, the subject-specific approach can be regarded as simply a "way to generate a covariance model" and thus a competitor to specifying a model for $\boldsymbol{\Sigma}_i$ directly as we did in Chapter 8. Accordingly, we can compare the population-averaged models like (1) with different choices for $\boldsymbol{\Sigma}_i$, as we considered in Homework 3, to subject-specific models with different assumptions on the $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$, and use AIC and BIC to determine a suitable overall model.

With this is mind, fit the subject-specific model in (b) under the four assumptions considered by the investigators in Problem 3(b) of Homework 4. Make a table of the AIC and BIC values from each of these fits. Compare these AIC and BIC values to those you obtained in Problem 3 of Homework 3. Based on these results and those from Homework 3, which model would you prefer?

2. A common way of treating patients with cardiovascular disease is by surgical intervention. In particular, such patients may arrive at a hospital with symptoms such as unstable angina or suspected myocardial infarction (heart attack), requiring that physicians perform an invasive procedure called a percutaneous coronary intervention (PCI) to investigate the extent to which coronary arteries might be blocked. During such an investigation, the blockage may be treated using a balloon to dislodge the blockage and widen the artery ("balloon angioplasty"); in addition, a device known as a stent may be inserted to prop the artery open.

   When such PCI procedures are performed, it is necessary for the subject to be treated with a drug that inhibits the aggregation of platelets in the blood. Informally, platelets are a blood constituent involved in clotting of the blood; clotting occurs when the platelets aggregate together in "clumps." To ensure that clotting does not interfere with the procedure, inhibition of the clotting mechanism is desirable; clotting during the procedure can lead to complications such as stroke or heart attack. A long-standing issue has been to determine which of two popular drugs elicits the most desirable pattern of inhibition of platelet aggregation.

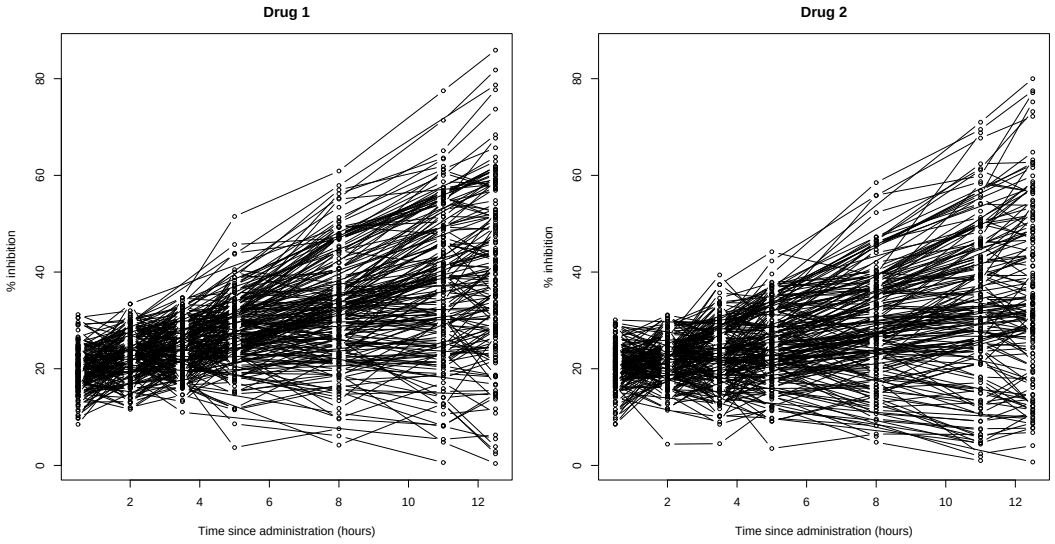
   Accordingly, an experiment was conducted to compare the platelet aggregation patterns of the drugs in such subjects under controlled conditions. Subjects arriving at a major medical center with symptoms of unstable angina or myocardial infarction who were judged to require a PCI procedure were randomized into two groups, one for each of the drugs, with 200 subjects per group. For each subject, at time 0, the assigned drug was administered according to the manufacturer's recommended dosage; for each drug, this involved giving the subject a large dose by injection to start inhibition of platelet aggregation immediately and simultaneously giving the subject a smaller dose of the drug intravenously at a constant rate over several hours, a method of administration known as an infusion. The purpose of the infusion was to

keep platelet aggregation inhibited over at least a 12 hour period, so that clotting would be minimized during the PCI procedure and subsequent recovery for the subject.

For each subject, blood samples were to be taken at 0.5, 2.0, 3.5, 5.0, 8.0, 11.0, and 12.5 hours. Each sample was to be analyzed for degree of platelet inhibition, characterized by the response "percent inhibition," a value between 0 and 100 representing the percentage of inhibition relative to that of an untreated sample (in units of "%μM"). Also recorded for each subject was whether the subject had experienced a previous myocardial infarction before the current hospitalization (0=no, 1=yes) and gender (0=female, 1=male).

The data from the study may be found on the class web page in the file `platelet.dat`. Each record in the file corresponds to a single observation, and the columns are (1) subject id number (1–400), (2) previous myocardial infarction indicator (0 = no, 1 = yes), (3) gender indicator (0 = female, 1 = male), (4) time (hours, measured since administration of drug), (5) percent inhibition, and (6) drug group indicator (1 or 2). Note that for some subjects, the response is not available at all intended time points; some samples were mishandled and in some instances study personnel did not follow the instructions and neglected to obtain samples. It was determined that the reasons for the missing values had nothing to do with the drugs or the patterns of inhibition. The data are depicted graphically in Figure 1.

Figure 1: *Percent platelet inhibition for two groups of subjects with cardiovascular disease*



From the plot, it appears that over the period of the study, platelet inhibition appears for most subjects to follow a rough straight-line trajectory that either stays relatively flat or rises, although a few profiles seem to decrease. To represent this, the investigators proposed the following model. Because the investigators were particularly interested in the time point 0.5 hours post-administration, as we will see shortly, they defined time in the model so that $t = 0$ corresponds to 0.5 hours after administration of the drug. That is, they let t_{ij}, the time of the jth platelet inhibition response on subject i, be defined as

$$t_{ij} = s_{ij} - 0.5,$$

where s_{ij} = time of the jth response on subject i measured from administration of the drug

(so s_{ij} equals the time value given in the data file). Letting y_{ij} be the corresponding platelet inhibition response for subject i at the jth time, the model is

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}, \tag{4}$$

where the parameters β_{0i} and β_{1i} describe the percent inhibition trajectory starting at 0.5 hours following drug administration ($t_{ij} = 0$) for the ith subject, and e_{ij} represents a mean-zero deviation associated with the jth inhibition response, assumed to be normally distributed. This model thus allows the pattern after 0.5 hours to follow a straight line for each subject.

(a) The investigators initially wished to assume that, for model (4), mean platelet inhibition at 0.5 hours following administration of drug has the following features within each drug group:

- is associated with whether the subject has had a previous myocardial infarction
- is associated with whether the subject is male or female
- the way in which it is associated with whether the subject is male or female is the different depending on whether the subject has had a previous myocardial infarction.

Because the subjects had been on the drugs for 0.5 hours, the investigators assumed that mean platelet inhibition at 0.5 hours and the way the above features occur is *different* for the two drugs.

The investigators also wished to assume that the "typical" or mean rate of change of platelet inhibition over the study period *also* has these features, and they wished to allow for the possibility that the mean rate of change of platelet inhibition and its association with prior myocardial infarction and gender could be different for each drug. This would allow the possibility that the drug that is received is associated with the pattern of change of platelet inhibition in different ways for subjects of different genders and prior history of myocardial infarction.

Let $m_i = 0$ if subject i has not had a previous myocardial infarction and $m_i = 1$ if s/he has, and let $g_i = 0$ if i is female, and $g_i = 1$ if i is male. Given these beliefs, write down expressions for β_{0i} and β_{1i} for subject i taking drug k, $k = 1, 2$. Be sure to define and fully describe all additional symbols you use.

(b) In terms of your model in (a)

(i) Give an expression that represents the typical value of platelet inhibition at 0.5 hours after drug administration for male subjects with a previous myocardial infarction taking drug 2.

(ii) Give an expression for mean platelet inhibition for female subjects with no previous myocardial infarction at 12 hours following administration of drug 1.

In the following parts, you will develop a SAS program to carry out several different analyses. You will have to modify your program for each part to obtain desired analyses. Please turn in your final program and output that carries out all necessary analyses.

(c) The investigators were all willing to believe that

(i) the assay used to measure platelet inhibition for both drug groups exhibits constant variation regardless of the true value of platelet inhibition being ascertained,

(ii) within-subject local "fluctuations" in platelet inhibition are of similar magnitude for both drugs and across time for all subjects,

(iii) variation in "inherent," true platelet inhibition at 0.5 hours is similar for patients in both drug groups, as is variation in the "inherent" rates of change of platelet inhibition over the study period and the way these quantities co-vary.

One of the investigators was concerned, however, that the time points at which platelet inhibition was measured were not sufficiently far apart in time to ensure that measurements within a subject are uncorrelated. He was willing to believe that, if such correlation is present, it "falls off" as the time points get farther apart, but he insisted that an analysis be done to resolve this issue.

Give two different sets of assumptions on the e_{ij}, $= 1, \ldots, n_i$, in 4) and random effects corresponding to β_{0i} and β_{1i} in (4) that incorporate (i)–(iii). The first set of assumptions should incorporate the investigator's concern; the second set should represent the case where the investigator's concern is unwarranted.

Fit the overall model (4) along with your model for β_{0i} and β_{1i} in (a) under both sets of assumptions using SAS `proc mixed` and the method of maximum likelihood. Which set of assumptions is best supported?

(d) From the output for the fit of the model you preferred in (c), write down an estimate of the variance associated with among-subject variation in true platelet inhibition in the population of male subjects with no previous myocardial infarction receiving drug 2 at 0.5 hours post-administration.

(e) Previous research has suggested that the way in which platelet inhibition occurs for both drugs over this period may be associated with whether a subject has had a previous myocardial infarction, but there is no evidence to suggest that it is associated with gender in any way. Thus, the investigators planned to base their subsequent analyses not on the model you developed in (a) but on a model that includes no effect of gender either in the representation of mean platelet inhibition at 0.5 hours or in the representation of the "typical" rate of change of platelet inhibition over the study period. Write down this simpler model and fit it using ML and your preferred covariance structure from (c). Based on your preferred fit in (c) and this fit, is there any evidence against doing this?

(f) *For the rest of the problem, consider the simpler model in (e) with no gender effects.* The reason that the investigators were so interested in 0.5 hours post-administration is because another research team had recently published a paper receiving a lot of press, which claimed that the 2 drugs exhibit the same mean platelet inhibition and that, furthermore, mean platelet inhibition on the two drugs is the same for subjects with or without a previous myocardial infarction. This team based their finding on comparing platelet inhibition levels 0.5 hours post-administration. Our investigators felt that comparing platelet inhibition at a single time point, particularly one so soon after administration, was not very informative. Thus, their first goal was to examine whether the data from the current study offer evidence refuting the claim of their rival investigators.

Write down a set of hypotheses that addresses the issue of interest to the investigators in terms of the model in (e), and express your null hypothesis in terms of a linear function $\boldsymbol{L\beta}$, defining $\boldsymbol{L}$. Using Wald methods, carry out the test at level of significance 0.05 based on a REML fit. State your conclusion as a *meaningful sentence*.

(g) The investigators' second goal was to make the point that comparing platelet inhibition at a single point does not tell the whole story. Thus, regardless of how the test in (f) turned out, they wanted to investigate longer time periods and the rate of change of platelet inhibition over them. The first question along these lines was whether the way "typical" rate of change

differs between subjects who have had a previous myocardial infarction and those who have not is different for the two drugs.

Write down a set of hypotheses that addresses the issue of interest to the investigators in terms of the model in (e), and express your null hypothesis in terms of a linear function $\boldsymbol{L\beta}$, defining $\boldsymbol{L}$. Using Wald methods, carry out the test at level of significance 0.05 based on a REML fit. State your conclusion as a *meaningful sentence*.

(h) The second question was whether, averaged across patients with and without a previous myocardial infarction, "typical" rate of change of platelet inhibition was different for the two drugs. Write down a set of hypotheses that addresses the issue of interest to the investigators in terms of the model in (e), and express your null hypothesis in terms of a linear function $\boldsymbol{L\beta}$, defining $\boldsymbol{L}$. Using Wald methods, carry out the test at level of significance 0.05 based on a REML fit. State your conclusion as a *meaningful sentence*.

(i) Based on the model in (e), provide estimates (and associated standard errors) of mean platelet inhibition at 12.5 hours after administration of (a) drug 1 in subjects with previous myocardial infarction; and (b) drug 2 in subjects with no previous myocardial infarction.

3. Consider again the platelet inhibition study in Problem 2. Consider also the model you specified in part (e).

   As before, you will develop a SAS program to carry out several different analyses. You will have to modify your program for each part to obtain desired analyses. Please turn in your final program and output that carries out all necessary analyses.

   (a) Using SAS `proc mixed`, fit the model using *maximum likelihood*. This time, have your program print out *both* the estimates of the fixed parameters in $\boldsymbol{\beta}$ *and* the approximate best linear unbiased predictors (BLUPs) of the random effects $\boldsymbol{b}_i$.

   (b) In your program, create 4 data sets and have the program print out their contents:

   (i) A data set containing only the data for subject 4 in drug group 1

   (ii) A data set containing only the data for subject 100 in drug group 1 item[(iii)] A data set containing only the data for subject 230 in drug group 2

   This may be carried out by using the `where` statement to create a new data set as follows; e.g., for (i), if the variable `patient` is subject indicator and all the data are in the data set `alldata`

```
data patient1; set alldata;
     where patient=1;
run;
```

   The resulting data set will contain only the data on all variables where `patient=1` (so for the first subject).

   (c) From the output of your `proc mixed` run in (a), calculate *by hand* the approximate BLUPs for intercept and slope (β_{0i} and β_{1i}) for each of the subjects in (b).

   (d) Consider the data for each subject in (b). Suppose we wished to estimate the subject-specific intercept and slope for each subject based only on his/her data alone.

   (i) Do the assumptions on within-individual variation in your model support doing this using ordinary least squares? Explain.

   (ii) Is it possible to obtain a reliable estimate for each of these subjects based on his/her data alone? Explain.

(e) Using `proc reg` or `proc glm`, estimate the intercept and slope separately for each of the subjects in (b) for whom reliable estimates are possible using ordinary least squares. Compare these to the BLUPs you calculated in (c).

(f) Suppose the model is thought to hold *except* that it is assumed that *all* of the subject-to-subject variation among slopes in each group is taken to be a consequence of whether or not the subjects have had a prior myocardial infarction, so that, once this factor is taken into account in slope, there is *no* apparent biological among-subject variation. Fit this model using `proc mixed` and the method of maximum likelihood, printing out the estimates of the fixed parameters $\boldsymbol{\beta}$ and the approximate BLUPs for $\boldsymbol{b}_i$.

Informally, which model do you prefer, the original one or this one? Explain.

(g) One of the investigators would like to conduct a formal test of whether there is evidence that the original model is better than the one in (f). Do whatever you think is appropriate to address this question.