ST 790, Midterm Solutions
Spring 2017

Please sign the following pledge certifying that the work on this test is your own:

"I have neither given nor received aid on this test."

Signature: _____

Printed Name: _____

There are FOUR questions, each with multiple parts. For each part of each question, please write your answers in the space provided. If you need more space, continue on the back of the page and indicate clearly where on the back you have continued your answer. Scratch paper is available from the instructor; just ask.

You are allowed ONE (1) SHEET of NOTES (front and back). Calculators are NOT allowed (you will not need one). NOTHING should be on your desk but this test paper, your one page of notes, and any scratch paper given to you by the instructor.

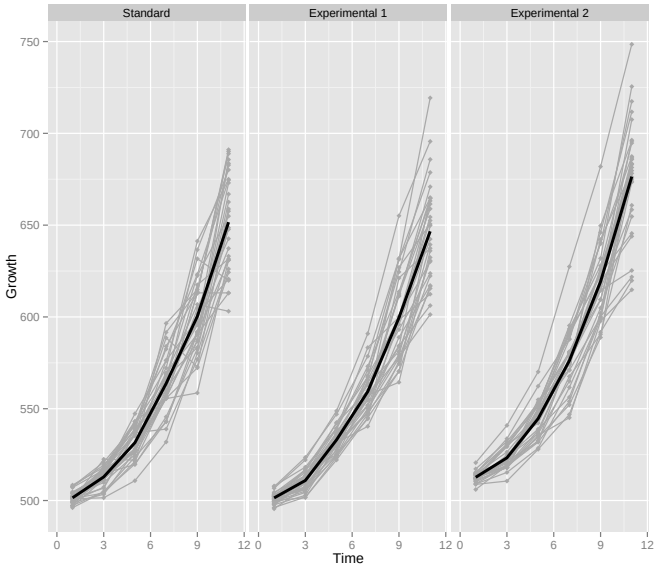
Points for each part of each problem are given in the left margin. TOTAL POINTS = 100.

If you are asked to provide an expression, you need not carry out the algebra to simplify the expression (unless you want to do so).

In all problems, all symbols and notation are defined exactly as they are in the class notes.

NOTE: My answers are MUCH MORE DETAILED than I expected yours to be.

1. A crop scientist has conducted an experiment to study the growth of three varieties of hops (*Humulus lupulus*, a plant used in the brewing of beer): a standard variety ("Standard") and two experimental varieties ("Experimental 1" and "Experimental 2"). 30 plants of each variety were planted at week 0, and at weeks 1, 3, 5, 7, 9, and 11, a measure of growth was obtained on each plant. The data are shown below, with the sample means at each measurement time superimposed.



The main goals of the experiment were

(i) To determine if the patterns of change of mean growth are not the same for all varieties.

(ii) To determine if the rate of change of mean growth is not constant for at least one of the varieties.

The crop scientist hopes to address this and other questions based on the following model and the standard assumptions made for it:

$$Y_{h\ell j} = \mu_{\ell j} + b_{h\ell} + e_{h\ell j} = \mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j} + b_{h\ell} + e_{h\ell j}, \qquad (1)$$

where $Y_{h\ell j}$ is growth for the hth plant from the ℓth variety at jth measurement occasion, $j = 1, \ldots, 6$; $\ell = 1, 2, 3$ indexes the Standard, Experimental 1, and Experimental 2 varieties, respectively; and the terms on the right hand side of (1) are as defined in the course notes.

The crop scientist presents you with the following output of an analysis based on (1):

Source	DF	Type III SS	Mean Square	F Value	Pr > F
variety	2	24512.63837	12256.31919	15.44	<.0001
Error	87	69070.26267	793.91107		
week	5	1423820.669	284764.134	1544.26	<.0001
week*variety	10	2154.105	215.410	1.17	0.3104
Error(week)	435	80214.903	184.402		

Mauchly's Criterion	DF	Chi-Square	Pr > ChiSq
	14	343.30055	<.0001

For each variety, the scientist has also calculated the sample covariance matrix $\widehat{\boldsymbol{V}}$ and associated correlation matrix $\widehat{\boldsymbol{\Gamma}}$ of the observed responses; these are as follows for the Standard (S), Experimental 1 ($E1$), and Experimental 2 ($E2$) varieties:

$$\widehat{\boldsymbol{V}}_S = \begin{pmatrix} 11.1 & 8.5 & 13.4 & 32.7 & 6.6 & 45.3 \\ 8.5 & 30.1 & 23.1 & 50.1 & 26.0 & 26.4 \\ 13.4 & 23.1 & 77.0 & 70.7 & 69.5 & 58.0 \\ 32.7 & 50.1 & 70.7 & 241.8 & 137.0 & 182.7 \\ 6.6 & 26.0 & 69.5 & 137.0 & 406.6 & 146.6 \\ 45.3 & 26.4 & 58.0 & 182.7 & 146.6 & 723.3 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}}_S = \begin{pmatrix} 1.00 & 0.47 & 0.46 & 0.63 & 0.10 & 0.50 \\ 0.47 & 1.00 & 0.48 & 0.59 & 0.23 & 0.18 \\ 0.46 & 0.48 & 1.00 & 0.52 & 0.39 & 0.25 \\ 0.63 & 0.59 & 0.52 & 1.00 & 0.44 & 0.44 \\ 0.10 & 0.23 & 0.39 & 0.44 & 1.00 & 0.27 \\ 0.50 & 0.18 & 0.25 & 0.44 & 0.27 & 1.00 \end{pmatrix}$$

$$\widehat{\boldsymbol{V}}_{E1} = \begin{pmatrix} 8.9 & 13.1 & 13.6 & 26.7 & 44.9 & 67.6 \\ 13.1 & 31.1 & 29.2 & 51.2 & 96.4 & 112.6 \\ 13.6 & 29.2 & 47.2 & 58.4 & 88.6 & 109.3 \\ 26.7 & 51.2 & 58.4 & 142.6 & 170.6 & 223.8 \\ 44.9 & 96.4 & 88.6 & 170.6 & 522.3 & 464.7 \\ 67.6 & 112.6 & 109.3 & 223.8 & 464.7 & 768.8 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}}_{E1} = \begin{pmatrix} 1.00 & 0.79 & 0.67 & 0.75 & 0.66 & 0.82 \\ 0.79 & 1.00 & 0.76 & 0.77 & 0.76 & 0.73 \\ 0.67 & 0.76 & 1.00 & 0.71 & 0.56 & 0.57 \\ 0.75 & 0.77 & 0.71 & 1.00 & 0.63 & 0.68 \\ 0.66 & 0.76 & 0.56 & 0.63 & 1.00 & 0.73 \\ 0.82 & 0.73 & 0.57 & 0.68 & 0.73 & 1.00 \end{pmatrix}$$

$$\widehat{\boldsymbol{V}}_{E2} = \begin{pmatrix} 7.7 & 12.4 & 21.4 & 35.2 & 42.2 & 44.6 \\ 12.4 & 41.7 & 47.3 & 79.7 & 81.5 & 150.4 \\ 21.4 & 47.3 & 99.1 & 133.6 & 167.7 & 223.7 \\ 35.2 & 79.7 & 133.6 & 279.5 & 281.2 & 353.6 \\ 42.2 & 81.5 & 167.7 & 281.2 & 422.7 & 421.2 \\ 44.6 & 150.4 & 223.7 & 353.6 & 421.2 & 993.4 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}}_{E2} = \begin{pmatrix} 1.00 & 0.69 & 0.77 & 0.76 & 0.74 & 0.51 \\ 0.69 & 1.00 & 0.74 & 0.74 & 0.61 & 0.74 \\ 0.77 & 0.74 & 1.00 & 0.80 & 0.82 & 0.71 \\ 0.76 & 0.74 & 0.80 & 1.00 & 0.82 & 0.67 \\ 0.74 & 0.61 & 0.82 & 0.82 & 1.00 & 0.65 \\ 0.51 & 0.74 & 0.71 & 0.67 & 0.65 & 1.00 \end{pmatrix}$$

[7 points] (a) Define

$$\mathcal{M} = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{16} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{26} \\ \mu_{31} & \mu_{32} & \cdots & \mu_{36} \end{pmatrix}.$$

Give an expression in terms of $\mathcal{M}$ that formalizes the crop scientist's question (ii) (to determine if the rate of change of mean growth is not constant for at least one of the varieties), defining any additional symbols you use, or explain why this is not possible.

This is a difficult question and rather mean of me to give to you. The plot of the sample means provides informal evidence that there may be curvature, and thus nonconstant rate of change of mean growth, for all three varieties. Unlike with our nice regression-like models, where we can represent this phenomenon directly and explicitly to address the question, model (1) does not allow this. So if we want to address this question with (1), we must find an indirect way to represent the null hypothesis that all three mean profiles show constant rate of change in terms of the means for each variety at each time.

Several of you said that this could be addressed by the usual test of *parallelism* (`week*variety` interaction); however, this is not the case. This hypothesis asks if the pattern of change is different across varieties, but the pattern can be different even if the rate of change is constant for all three varieties. To see this, think of the situation where the mean response profiles are straight lines, so do exhibit constant rate of change for all three varieties, but have very different slopes. This test would reject the null hypothesis of parallelism in a situation where the rate of change is in fact constant for all three varieties.

Here, the time points are *equally spaced*. Note that under this condition, if for each group $\ell = 1, 2, 3$

$$\mu_{\ell 2} - \mu_{\ell 1} = \mu_{\ell 3} - \mu_{\ell 2} = \mu_{\ell 4} - \mu_{\ell 3} = \mu_{\ell 5} - \mu_{\ell 4} = \mu_{\ell 6} - \mu_{\ell 5}, \tag{2}$$

or, equivalently,

$$\mu_{\ell 1} - \mu_{\ell 2} = \mu_{\ell 2} - \mu_{\ell 3} = \mu_{\ell 3} - \mu_{\ell 4} = \mu_{\ell 4} - \mu_{\ell 5} = \mu_{\ell 5} - \mu_{\ell 6},$$

then the $\mu_{\ell j}$ lie on a straight line, so that the rate of change of mean growth is indeed constant. Thus, the null hypothesis that the rate of change is constant for all varieties can be represented by (2) for $\ell = 1, 2, 3$, with the alternative that this is not true for at least one of the varieties (so that the rate of change is not constant). I was happy if you wrote this out (it is in terms of elements of $\mathcal{M}$).

It is possible to represent the null hypothesis several different ways in terms of $\mathcal{M}$. One quick way is to notice that (2) can be written as

$$\begin{aligned}
\mu_{\ell 2} - \mu_{\ell 1} - (\mu_{\ell 3} - \mu_{\ell 2}) &= 0 \\
\mu_{\ell 4} - \mu_{\ell 3} - (\mu_{\ell 3} - \mu_{\ell 2}) &= 0 \\
\mu_{\ell 5} - \mu_{\ell 4} - (\mu_{\ell 4} - \mu_{\ell 3}) &= 0 \\
\mu_{\ell 6} - \mu_{\ell 5} - (\mu_{\ell 5} - \mu_{\ell 4}) &= 0
\end{aligned}$$

or equivalently

$$\begin{aligned}
\mu_{\ell 1} - 2\mu_{\ell 2} + \mu_{\ell 3} &= 0 \\
\mu_{\ell 2} - 2\mu_{\ell 3} + \mu_{\ell 4} &= 0 \\
\mu_{\ell 3} - 2\mu_{\ell 4} + \mu_{\ell 5} &= 0 \\
\mu_{\ell 4} - 2\mu_{\ell 5} + \mu_{\ell 6} &= 0
\end{aligned}$$

for $\ell = 1, 2, 3$, which can be gotten by post-multiplying $\mathcal{M}$ by an appropriate $\boldsymbol{U}$ matrix directly such a $\boldsymbol{U}$ is

$$\boldsymbol{U} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Thus, with this $\boldsymbol{U}$, the null hypothesis can be written as $\mathcal{M}\boldsymbol{U} = \boldsymbol{0}$.

Alternatively, we can take advantage of the fact we know that the successive differences in means can be gotten in a (6×5) by post-multiplying $\mathcal{M}$ as usual by

$$\boldsymbol{U} = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(or the negative of this matrix). We then get as usual

$$\mathcal{M}\boldsymbol{U} = \begin{pmatrix} \mu_{12} - \mu_{11} & \mu_{13} - \mu_{12} & \mu_{14} - \mu_{13} & \mu_{15} - \mu_{14} & \mu_{16} - \mu_{15} \\ \mu_{22} - \mu_{21} & \mu_{23} - \mu_{22} & \mu_{24} - \mu_{23} & \mu_{25} - \mu_{24} & \mu_{26} - \mu_{25} \\ \mu_{32} - \mu_{31} & \mu_{33} - \mu_{32} & \mu_{34} - \mu_{33} & \mu_{35} - \mu_{34} & \mu_{36} - \mu_{35} \end{pmatrix}. \tag{3}$$

To get (2) for each group, we want to get the contrasts of the successive differences in each row of $\mathcal{M}\boldsymbol{U}$. This can be accomplished by simply post-multiplying $\mathcal{M}\boldsymbol{U}$ by

$$\boldsymbol{U}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

or other matrix that takes differences. The null hypothesis can then be expressed as $\mathcal{M}\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{0}$.

The above is a formal solution to this problem. Another, more informal approach would be to consider the post-multiplying matrix $\boldsymbol{U}$ whose columns are orthogonal polynomial contrasts and to inspect the tests for quadratic and higher order components. If, for example, there is evidence of a difference in quadratic components across varieties, this might be reflecting that fact that the mean profile for at least one variety exhibits curvature, suggesting that the rate of change is not constant. Alternatively, if there is evidence of a main quadratic effect of time (so averaged across varieties), this could also be interpreted to be reflecting that at least one variety exhibits curvature. But these tactics are only informal and do not address the issue explicitly as in (2).

Some of you said that this is not possible. I don't blame you; we did discuss the fact that the classical approach does not allow one to, for example, *characterize* and estimate rates of change explicitly the way a regression-type model does. However, where we are interested in testing a hypothesis about rates of change, even though we can't estimate these, we can represent what would have to hold for the rate of change to be constant for all varieties (the null hypothesis of interest) in terms of the elements of $\mathcal{M}$. Here, with equally spaced time points, this is easy. With unequally spaced time points, we would need to incorporate the time points, and this would be more of a mess.

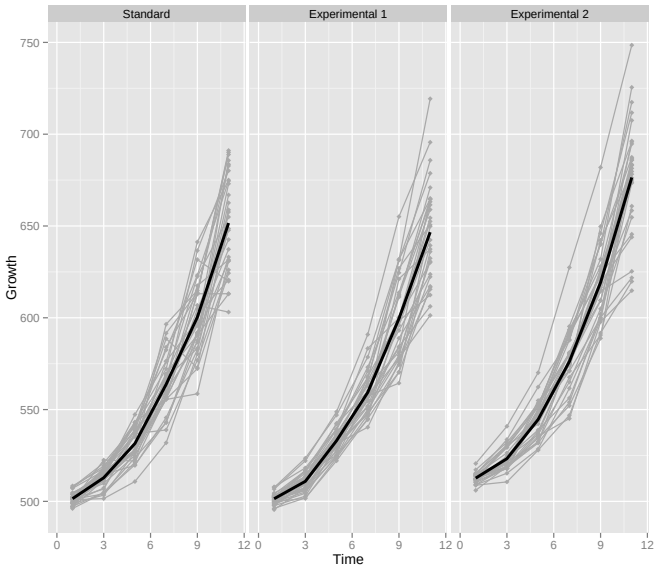
[10 points] (b) Based on the information you have, can you address question *reliably* (i) using model (1)? If so, describe how and present a formal statement of the result. If not, explain why not.

The best answer is no, because it appears that the assumptions required to ensure valid inferences are likely violated.

If all of the assumptions required to ensure validity of tests based on (1) were satisfied, we would address (i) by the test of parallelism, which in the ANOVA table is given in the `week*variety` row. These assumptions include: (i) the covariance matrix of a data vector is the same for all individuals, regardless of group; and (ii) the common covariance matrix is compound symmetric with the same variance at all time points (although this can be relaxed to a common Type H structure).

The assumption of a common structure is questionable based on the sample covariance/correlation matrices, which do not all exhibit similar correlation patterns. Even if the sample correlation pattern is "similar enough" to be willing to say it might be the same, the sample covariance matrices clearly indicate that variance likely increases with time. This might be possible under a more general Type H structure, but even if the correlation structure were the same for all three varieties, the test of sphericity under that assumption (Mauchly's Criterion) strongly rejects the null hypothesis that the structure is of Type H. In short, the evidence available does not seem to support the key assumptions needed.

2. Consider the hops experiment in the previous problem. Here are the data again:



The goals are

(i) To determine if the patterns of change of mean growth are not the same for all varieties.

(ii) To determine if the rate of change of mean growth is not constant for at least one of the varieties.

[10 points] (a) Propose a statistical model different from that in (1) in which both (i) and (ii) can be addressed. *Briefly* state any assumptions you incorporate in the model.

You needed to write down a model and define all the components.

The questions of interest are clearly *population-averaged* questions. It is equally valid to posit a PA linear model directly or to posit a SS linear mixed effects model to induce a PA model indirectly. Given we have sample information on the overall covariance/correlation structure for each variety, I show a direct PA model here. If you used a linear mixed effects model to induce a PA model, which some of you did, that's fine, too.

From the plots, the mean response in each group appears to be reasonably approximated by a quadratic function in time. Letting Y_{ij} be response at the jth week t_j on the ith plot, $j = 1, \ldots, 6$, $i = 1, \ldots, 90$, $\delta_{i\ell} = 1$ if i is in variety ℓ and = 0 otherwise, $\ell = 1, 2, 3$, a plausible model is

$$Y_{ij} = (\beta_{01}\delta_{i1} + \beta_{02}\delta_{i2} + \beta_{03}\delta_{13}) + (\beta_{11}\delta_{i1} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{13})t_j + (\beta_{21}\delta_{i1} + \beta_{22}\delta_{i2} + \beta_{23}\delta_{13})t_j^2 + \epsilon_{ij}.$$

We have allowed the intercept to be different for each variety in accordance with the visual evidence in the plot of the data, which suggests that at the very least the mean response at baseline may be higher for Experimental 2 than the other two treatments; and $\boldsymbol{a}_i = (\delta_{i1}, \delta_{i2}, \delta_{i3})^T$. Some of you took the intercept to be the same for all varieties; this really is not consistent with the evidence. A few of you did this on the basis of randomization; however, here, the plants were not randomized to be of different varieties but were instead randomly chosen from the populations of Standard, Experiemental 1, and Experimental 2 varieties. So there is no reason to expect the mean growth responses to be the same on this basis.

Based on the sample information above, we assume $E(\epsilon_{ij}|\boldsymbol{a}_i) = 0$, and, with $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{i6})^T$,

$$\text{var}(\boldsymbol{\epsilon}_i|\boldsymbol{a}_i) = \boldsymbol{V}_\ell = \boldsymbol{T}_\ell^{1/2}\boldsymbol{\Gamma}_\ell\boldsymbol{T}_\ell^{1/2},$$

where $\ell = 1, 2$, or 3 depending on $\boldsymbol{a}_i$. Here, ϵ_{ij} represents the aggregate deviation from the mean growth at week$j$ for Y_{ij} due to among- and within-individual sources. Given the sample information, it might be reasonable to assume possibly different covariance matrices for each variety as above. In any case, the evidence from the sample covariance matrices in Problem 1 definitely suggests the possibility that variance changes over time for each variety, so we might assume $\text{var}(\epsilon_{ij}|\boldsymbol{a}_i) = \sigma^2_{\ell j}$, so that $\boldsymbol{T}_\ell = \text{diag}(\sigma^2_{\ell 1}, \ldots, \sigma^2_{\ell 6})$. We might also take $\boldsymbol{\Gamma}_\ell$ to be compound symmetric for each variety based on the sample evidence, with different correlation parameter α_ℓ for each variety. Another option is to just take $\boldsymbol{V}_\ell$ to be completely unstructured for each variety (which implies variances that are not constant for all times). We might also be willing to make the assumption that $Y_i|\boldsymbol{a}_i$ under these conditions is normal with moments implied by these specifications, although that is not absolutely necessary.

[8 points] (b) For your model in (a), write down a vector $\boldsymbol{\beta}$ that collects all parameters that characterize mean growth for the three varieties. Then provide a matrix $\boldsymbol{L}$ such that you can address the question of whether or not the rate of change in mean growth is *both* constant and the same for all varieties through an expression of the form $\boldsymbol{L}\boldsymbol{\beta}$.

$$\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23})^T.$$

Under the above model, the rate of change of mean growth at any time t is, taking derivatives with respect to time,

$$(\beta_{11}\delta_{i1} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{13}) + 2(\beta_{21}\delta_{i1} + \beta_{22}\delta_{i2} + \beta_{23}\delta_{13})t.$$

If this is to be constant for all varieties, it must be that $\beta_{21} = \beta_{22} = \beta_{23} = 0$; if it is furthermore the same for each variety, it must be that $\beta_{11} = \beta_{12} = \beta_{13}$. This leads to

$$\boldsymbol{L} = \begin{pmatrix} 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

If you parameterized your model differently, I graded your $\boldsymbol{L}$ according to your model.

[8 points] (c) In terms of $\boldsymbol{\beta}$ you defined in (b), provide a matrix $\boldsymbol{L}$ that allows you to characterize the rate of change of mean growth for each variety at the midpoint of the study period (5.5 weeks) through an expression of the form $\boldsymbol{L\beta}$.

As above, the rate of change at any time t for variety ℓ is the derivative of the population mean with respect to t,

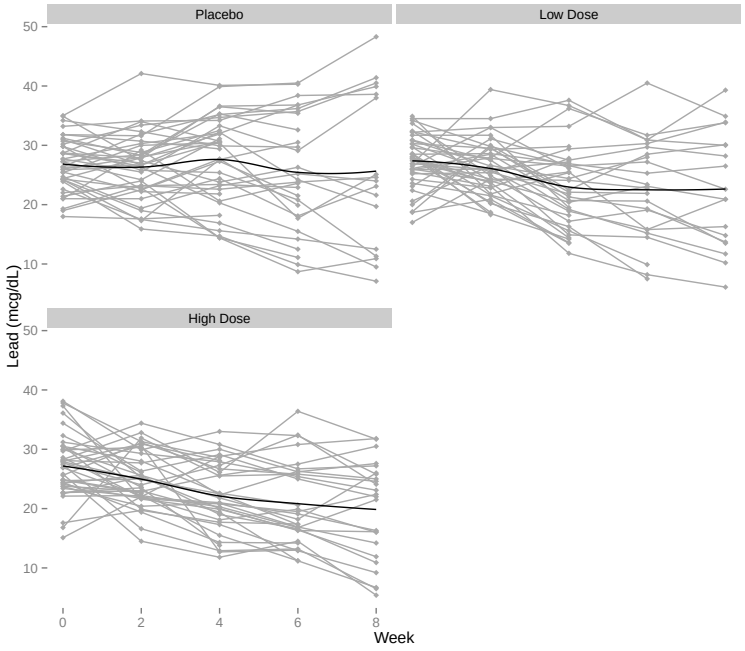
$$\beta_{1\ell} + 2\beta_{2\ell}t.$$

Thus, if we set $t = 5.5$, we get the rate of change at the midpoint of the study. We can get all three rates simultaneously using $2 \times 5.5 = 11$ and

$$\boldsymbol{L} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 11 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 11 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 11 \end{pmatrix}.$$

If you parameterized your model differently, I of course graded your answer accordingly.

3. The data shown shown below are from a study of an oral treatment for lead exposure in 120 children with blood lead levels greater than 10 μg/dL. The children were randomized to receive placebo, low dose of the active treatment, or high dose of the active treatment, 40 children per group. Blood lead levels were to be obtained from each child at baseline (week 0), prior to initiation of assigned treatment, and then at weeks 2, 4, 6, 8 thereafter. The gender of each child was also recorded (0 = female, 1 = male). Here are the data, with a loess smooth superimposed on each plot.



There was substantial dropout from the study; only 84 of the 120 children returned for the week 6 measurement, and only 59 returned at week 8.

One of the study investigators ("Investigator A") posed the following questions:

(i) Is the average (mean) rate of change in lead levels over the 8 weeks associated with treatment received?

(ii) Is the average rate of change in lead levels over the 8 weeks associated with gender?

(iii) Is average baseline lead level different for males and females?

Another investigator ("Investigator B") disagreed, and restated the questions as follows:

(i) Is the rate of change of average (mean) lead level over the 8 weeks associated with treatment received?

(ii) Is the rate of change in average lead level over the 8 weeks associated with gender?

(iii) Is the average baseline lead level different for males and females?

[12 points] (a) Can you propose a statistical model in which questions (i)-(iii) of both Investigator A and Investigator B can be addressed? If so, write down the model and *briefly* state any assumptions you incorporate in the model. If not, state why not, and write down a model in which the questions of *one of* the investigators can be addressed (state which investigator), including (*briefly*) any assumptions you incorporate in the model.

The short answer is: Yes, you can. Investigator A's questions are SS questions, while those of Investigator B are PA questions. Thus, a linear mixed effects model is the obvious choice, as it is a SS model that induces a PA model, so that the interpretation of the parameters in the implied population mean model can be either SS or PA.

I expected you to write out a basic model and briefly define the major components. Here is a detailed description of a possible model; I did not expect you to take the time to define things precisely the way I do here.

Letting Y_{ij} be lead level measured on child i at the jth time t_{ij} $j = 1, \ldots, n_i$ (possibly different for each i due to dropout), it is natural from the plot to model child-specific trajectories as straight lines, i.e., take the child-level model to be

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}, \qquad \boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i})^T,$$

which we could also write as $\boldsymbol{Y}_i = \boldsymbol{C}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i$ as in the notes. Define $\delta_{i\ell} = 1$ if i was randomized to treatment ℓ and = 0 otherwise, $\ell = 1, 2, 3$, where placebo=1, low dose=2, high dose=3, and let $g_i = 0$ if i is a girl and $g_i = 1$ if i is a boy. A population model that allows the above questions to be addressed is

$$\begin{aligned} \beta_{0i} &= \beta_{00} + \beta_{01} g_i + b_{0i} \\ \beta_{1i} &= (\beta_{11} + \beta_{21} g_i) + (\beta_{12} + \beta_{22} g_i)\delta_{i2} + (\beta_{13} + \beta_{23} g_i)\delta_{i3} + b_{1i}, \end{aligned} \tag{4}$$

where $\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$ is the vector of individual-specific random effects. We could equally well have parameterized this instead as

$$\begin{aligned} \beta_{0i} &= \beta_{00}(1 - g_i) + \beta_{01} g_i + b_{0i} \\ \beta_{1i} &= \{\beta_{11}(1 - g_i) + \beta_{21} g_i\} + \{\beta_{12}(1 - g_i) + \beta_{22} g_i\}\delta_{i2} + \{\beta_{13}(1 - g_i) + \beta_{23} g_i\}\delta_{i3} + b_{1i}, \end{aligned}$$

These specifications of β_{1i} allow the association of average rate of change (from a SS perspective) to be associated with gender in a way that depends on treatment; an example of a simpler such model is

$$\beta_{1i} = \beta_{11} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3} + \beta_2 g_i + b_{1i}. \tag{5}$$

You may have used a fancier or simpler model and parameterized it as above or differently.

Letting $\boldsymbol{a}_i = (\delta_i, g_i)^T$, we need to make assumptions on $\boldsymbol{e}_i$ and $\boldsymbol{b}_i$, for which we definitely assume $E(\boldsymbol{e}_i|ba_i) = \mathbf{0}$ and $\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \mathbf{0}$. To complete the specification, in particular of the forms of $\text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \boldsymbol{R}_i(\gamma)$ and $\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i)$, it would be great to be able to see fits under different choices for these (and the associated AIC and BIC values) and to see residual plots, which would also give us an informal assessment of whether or not $\boldsymbol{e}_i|\boldsymbol{a}_i$ and $\boldsymbol{b}_i|\boldsymbol{a}_i$ are approximately normal. The default specification is of course that $\text{var}(\boldsymbol{e}_i|\boldsymbol{a}_i) = \sigma^2\boldsymbol{I}_{n_i}$ and $\text{var}(\boldsymbol{b}_i|\boldsymbol{a}_i) = \boldsymbol{D}$ (2×2), but these could be relaxed to allow σ^2 and $\boldsymbol{D}$ to differ by group and/or to allow $\text{var}(e_{ij}|\boldsymbol{a}_i)$ to change over time and for within-individual correlation.

You may have written down a different model, which for most of you was fine; the key is that your model allows all the questions of interest above to be addressed. Your solution was graded based on the model you proposed.

[7 points] (b) In terms of your model in (a), show how you would address Investigator A's question (i) (Is the average rate of change in lead levels over 8 weeks associated with treatment received?) If you cannot, state why not.

We show this in the model (4) where the average rate of change in lead levels is

$$(\beta_{11} + \beta_{21} g_i) + (\beta_{12} + \beta_{22} g_i)\delta_{i2} + (\beta_{13} + \beta_{23} g_i)\delta_{i3}.$$

This is not associated with treatment if $\beta_{12} = \beta_{22} = \beta_{13} = \beta_{23} = 0$. So one would address this by testing this null hypothesis against the alternative that at least one of these parameters is different from zero, in which case the average rate of change is associated with treatment for at least one gender. In the simpler model (5), the average rate of change is

$$\beta_{11} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3} + \beta_2 g_i$$

and is not associated with treatment if $\beta_{12} = \beta_{13} = 0$.

Your solution was graded depending on the model you proposed. Many of you correctly noted that there are missing data, so one should be willing to assume missing at random (MAR) and normality, use maximum likelihood and model-based standard errors (ideally based on the observed information matrix).

[7 points] (c) In terms of your model in (a), show how you would address Investigator B's question (ii) (Is the rate of change in average lead level associated with gender?) If you cannot, state why not.

Because this is a linear model, the rate of change in average lead level is equal to the average rate of change, which in the fancier model (4) is again

$$(\beta_{11} + \beta_{21} g_i) + (\beta_{12} + \beta_{22} g_i)\delta_{i2} + (\beta_{13} + \beta_{23} g_i)\delta_{i3}.$$

This is not associated with gender if $\beta_{21} = \beta_{22} = \beta_{23} = 0$. If you opted for the simpler model (5), this boils down to $\beta_2 = 0$. Your solution was graded depending on the model you proposed. Again, many of you correctly made the point regarding missing data as in (b).

[8 points] (d) Show how, using your model in (a), you would estimate lead level at week 6 for the *i*th child, a female who received high dose active treatment. If you can not, state why not.

Taking model (4) as an example, a natural "estimator" for lead level at week t for the ith child in general is

$$\widehat{\beta}_{00} + \widehat{\beta}_{01} g_i + \{(\widehat{\beta}_{11} + \widehat{\beta}_{21} g_i) + (\widehat{\beta}_{12} + \widehat{\beta}_{22} g_i)\delta_{i2} + (\widehat{\beta}_{13} + \widehat{\beta}_{23} g_i)\delta_{i3}\} t + \widehat{b}_{0i} + \widehat{b}_{1i} t,$$

where the $\widehat{\beta}$'s are the usual ML estimators for the fixed effects (given the missing data) and $\widehat{b}_{0i}$ and $\widehat{b}_{1i}$ are the EBLUPs/empirical Bayes estimators for the random effects. Here, $t = 6$, $g_i = 0$, $\delta_{i2} = 0$, $\delta_{i3} = 1$, so that the "estimator" is

$$\widehat{\beta}_{00} + (\widehat{\beta}_{11} + \widehat{\beta}_{13})6 + \widehat{b}_{0i} + \widehat{b}_{1i} 6.$$

A similar expression is obtained with the simpler model (5). Your solution was graded based on your proposed model.

[8 points] (e) Suppose that children who dropped out of the study tended to do so because their previously recorded lead levels were not improving. Would you feel comfortable proceeding with a standard analysis using your model? If so, explain *briefly* how you would conduct the analysis based on your model under this condition. If not, explain why not.

You probably answered yes. If we have confirmation that this is the case, then it may be plausible to assume that the missingness mechanism is missing at random (MAR). If we are *also* willing to assume that lead levels are normally distributed given covariates and that the covariance structure is correctly specified, then under this condition and MAR we know that the usual maximum likelihood analysis will lead to valid inferences as long as we use model-based standard errors, ideally calculated based on the observed information matrix; also, likelihood ratio tests comparing nested models (so looking at simplifications of the above model as would be the case if (i) and (ii) hold).

[8 points] 4. (a) Consider the model

$$Y_{ij} = \beta_{0i} + e_{ij}, \quad i = 1, \ldots, m, \;\; j = 1, \ldots, n, \quad \beta_{0i} = \beta_0 + b_i, \tag{6}$$

where b_i are independent for all i, e_{ij} are independent for all i, j, and b_i and e_{ij} are independent of one another for all i, j, with

$$E(b_i) = 0, \quad \text{var}(b_i) = D \qquad E(e_{ij}) = 0, \quad \text{var}(e_{ij}) = \sigma^2.$$

Suppose we fit (6) by maximum likelihood to data under the assumption that b_i and e_{ij} are all normally distributed using standard software to obtain values for the estimators $\widehat{\beta}_0$, $\widehat{D}$, and $\widehat{\sigma}^2$ and associated standard errors for all three parameters based on model (6).

What conditions would you want to be met to feel comfortable that these standard errors for $\widehat{D}$ and $\widehat{\sigma}^2$ are reliable assessments of the uncertainty of estimation of D and σ^2? Explain (*briefly*) your answer.

The short answer is: The key conditions are (i) normality, (ii) correct model, and (iii) m large enough. Conditions (ii) and (iii) are always needed if we are to feel confident using an asymptotic approximation to the sampling distribution of any estimator. For estimators of parameters like D and σ^2, as we saw in Problem 1 of Homework 3, characterizing covariance structure, (i) is critical.

As we saw in that problem, the standard errors for components of the covariance parameter $\boldsymbol{\xi}$, which equals (σ^2, D) here, in particular the leading "2" in the expression for the covariance matrix of $\widehat{\boldsymbol{\xi}}$, depend critically on the assumption of normality holding. As we discussed, standard software bases computation of these on the result you derived in that problem; i.e., assuming normality. If the distribution is not normal, then this expression does not faithfully represent the true sampling variation, which depends on the true third and fourth moments. The result could be invalide characterization of the uncertainty of estimation of these parameters.

Thus, more precisely, an ideal answer is that we would want the true distribution of $\boldsymbol{Y}_i$ (there are no covariates here) to be *multivariate normal* with mean $\beta_0 \mathbf{1}_n$ and covariance matrix $\boldsymbol{V} = \sigma^2 \boldsymbol{I}_n + D\boldsymbol{J}_n$, and for m to be "large enough" for the asymptotic theory approximation to be reasonable.

[7 points] (b) Now consider the model

$$Y_{ij} = \beta_{0i} \exp(-\beta_1 t_j) + e_{ij}, \quad i = 1, \ldots, m, \;\; j = 1, \ldots, n, \quad \beta_{0i} = \beta_0 + b_i, \tag{7}$$

for times $t_1, \ldots, t_n$, where b_i and e_{ij} are as in (a).

Statistician A refers to β_1 in (7) as the average rate of decay of the response in the population, while Statistician B refers to β_1 as the rate of decay of the population average response. Which statistician is correct? Explain (*briefly*) your answer.

The short answer is: *both* are correct, because the model (7) is *linear* in β_{0i} and thus b_i.

This is exactly the situation in Homework 1, Problem 2(d). We can write model (7) as

$$Y_{ij} = \beta_0 \exp(-\beta_1 t_j) + b_{0i} \exp(-\beta_1 t_j) + e_{ij}.$$

It follows that the implied population mean/average response is

$$\beta_0 \exp(-\beta_1 t_j),$$

so that β_1 can be interpreted as the rate of decay of the population average response, making Statistician B, who is taking a PA perspective, correct. It is equally valid to regard (7) as a SS model, where the individual-specific linear coefficient β_{0i} varies in the population, but the individual-specific rate of decay β_1 does not. From the perspective of the hierarchical modeling perspective in Chapter 6, think of (7) as

$$Y_{ij} = \beta_{0i} \exp(-\beta_{1i} t_j) + e_{ij},$$

where

$$\beta_{0i} = \beta_0 + b_{0i}, \qquad \beta_{1i} = \beta_1,$$

so that the individual-specific rate of decay β_{1i} has no associated random effect, reflecting no variation in of these rates in the population, so that $\beta_{1i} = \beta_1$ for all i. Thus, β_1 has the equally valid interpretation as the average rate of decay of the response in the population (where there is no variation of individual rates about the average in the population), and Statistician A is also correct.