ST 790, Midterm Solutions
Spring 2018

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, most 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 physician who specializes in weight loss management has conducted a study to evaluate three weight loss programs for men. The physician recruited 120 overweight men weighing at least 220 pounds whose target, healthy weight was below 200 pounds. Each man was randomly assigned to one of three weight loss programs, 40 men per program. The weight of each man was measured at baseline (month 0), and each man returned to the clinic at months 1, 2, 3, 4, and 5 to have his weight recorded.

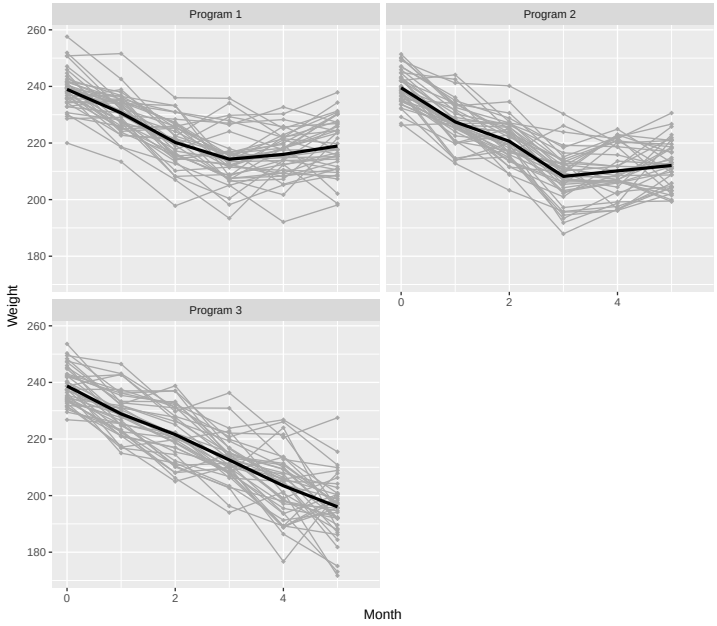
   The programs were as follows:

Program 1	Restricted diet
Program 2	Restricted diet plus weekly coaching on diet, exercise, and lifestyle
Program 3	Restricted diet and weekly coaching plus exercise with personal trainer 3 days/week

   For all three programs, meals on the restricted diet were provided to each man for the first 3 months of the study. Weekly coaching was also provided to each man assigned to Programs 2 and 3 for the first 3 months. Exercise sessions with the personal trainer were provided for the first 3 months to each man assigned to Program 3.

   For the last 2 months, all men were left on their own and received no meals, coaching, or exercise sessions.

   The data are shown below, with the sample means at baseline and at each monthly visit thereafter superimposed.



   The main goals of the study were

   (i) To determine if the pattern of change of mean weight is not the same for all programs over the 5 month study period.

   (ii) To determine if the rate of change of mean weight in the last 2 months is different from that in the first 3 months for at least one of the programs (suggesting that the effect of at least one of the weight loss programs is not sustained after it ends).

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

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

where $Y_{h\ell j}$ is the weight for the hth man assigned to the ℓth program at jth month, $j = 1, \ldots, 6$; $\ell = 1, 2, 3$ indexes Programs 1, 2, and 3, respectively; and the terms on the right hand side of (1) are as defined in the course notes.

Here are the sample covariance matrices $\widehat{\boldsymbol{V}}$ and associated correlation matrices $\widehat{\boldsymbol{\Gamma}}$ based on the data for each program (1, 2, 3):

$$\widehat{\boldsymbol{V}}_1 = \begin{pmatrix} 48.8 & 26.6 & 28.7 & 38.6 & 29.6 & 44.9 \\ 26.6 & 49.5 & 34.1 & 42.5 & 29.6 & 37.1 \\ 28.7 & 34.1 & 66.5 & 42.7 & 45.0 & 53.8 \\ 38.6 & 42.5 & 42.7 & 87.4 & 49.6 & 55.3 \\ 29.6 & 29.6 & 45.0 & 49.6 & 66.3 & 51.4 \\ 44.9 & 37.1 & 53.8 & 55.3 & 51.4 & 94.1 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}}_1 = \begin{pmatrix} 1.00 & 0.54 & 0.50 & 0.59 & 0.52 & 0.66 \\ 0.54 & 1.00 & 0.59 & 0.65 & 0.52 & 0.54 \\ 0.50 & 0.59 & 1.00 & 0.56 & 0.68 & 0.68 \\ 0.59 & 0.65 & 0.56 & 1.00 & 0.65 & 0.61 \\ 0.52 & 0.52 & 0.68 & 0.65 & 1.00 & 0.65 \\ 0.66 & 0.54 & 0.68 & 0.61 & 0.65 & 1.00 \end{pmatrix}$$

$$\widehat{\boldsymbol{V}}_2 = \begin{pmatrix} 37.6 & 25.8 & 26.1 & 33.0 & 36.6 & 24.3 \\ 25.8 & 52.9 & 29.5 & 37.6 & 35.7 & 26.0 \\ 26.1 & 29.5 & 53.7 & 43.5 & 40.5 & 32.6 \\ 33.0 & 37.6 & 43.5 & 84.7 & 59.0 & 31.1 \\ 36.6 & 35.7 & 40.5 & 59.0 & 77.8 & 34.5 \\ 24.3 & 26.0 & 32.6 & 31.1 & 34.5 & 68.9 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}}_2 = \begin{pmatrix} 1.00 & 0.58 & 0.58 & 0.58 & 0.68 & 0.48 \\ 0.58 & 1.00 & 0.55 & 0.56 & 0.56 & 0.43 \\ 0.58 & 0.55 & 1.00 & 0.64 & 0.63 & 0.54 \\ 0.58 & 0.56 & 0.64 & 1.00 & 0.73 & 0.41 \\ 0.68 & 0.56 & 0.63 & 0.73 & 1.00 & 0.47 \\ 0.48 & 0.43 & 0.54 & 0.41 & 0.47 & 1.00 \end{pmatrix}$$

$$\widehat{\boldsymbol{V}}_3 = \begin{pmatrix} 41.4 & 27.7 & 37.1 & 34.1 & 37.4 & 39.8 \\ 27.7 & 58.5 & 53.3 & 44.7 & 63.9 & 38.2 \\ 37.1 & 53.3 & 82.6 & 46.9 & 68.7 & 54.1 \\ 34.1 & 44.7 & 46.9 & 65.8 & 58.5 & 47.1 \\ 37.4 & 63.9 & 68.7 & 58.5 & 135.6 & 44.9 \\ 39.8 & 38.2 & 54.1 & 47.1 & 44.9 & 121.6 \end{pmatrix}, \quad \widehat{\boldsymbol{\Gamma}}_3 = \begin{pmatrix} 1.00 & 0.56 & 0.64 & 0.65 & 0.50 & 0.56 \\ 0.56 & 1.00 & 0.77 & 0.72 & 0.72 & 0.45 \\ 0.64 & 0.77 & 1.00 & 0.64 & 0.65 & 0.54 \\ 0.65 & 0.72 & 0.64 & 1.00 & 0.62 & 0.53 \\ 0.50 & 0.72 & 0.65 & 0.62 & 1.00 & 0.35 \\ 0.56 & 0.45 & 0.54 & 0.53 & 0.35 & 1.00 \end{pmatrix}$$

And here is the output of an analysis based on (1):

Source	DF	Type III SS	Mean Square	F Value	Pr > F
program	2	4766.50353	2383.25176	8.69	0.0003
Error	117	32088.16975	274.25786		
month	5	88497.18261	17699.43652	563.83	<.0001
month*program	10	10458.74281	1045.87428	33.32	<.0001
Error(month)	585	18363.92125	31.39132		

Mauchly's Criterion	DF	Chi-Square	Pr > ChiSq
	14	39.213961	0.0003

Define

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

[5 points] (a) Give an expression in terms of $\mathcal{M}$ that can be viewed as formalizing the physician's question (i), to determine if the pattern of change of mean weight is not the same for all programs over the 5 month study period, defining any additional symbols you use, or explain why this is not possible.

All of you correctly recognized that this question could be formalized by considering the test for parallelism of the form $H_0 : \boldsymbol{C}\mathcal{M}\boldsymbol{U} = \boldsymbol{0}$, where each of the matrices $\boldsymbol{C}$ and $\boldsymbol{U}$ is in the form of a "differencing" matrix. The most straightforward choices are

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

the so-called profile transformation matrix; some of you used the Helmert tranformation matrix instead, which is fine.

[8 points] (b) Given an expression in terms of $\mathcal{M}$ that can be viewed as formalizing the physician's question (ii), to determine if the rate of change of mean weight in the last 2 months is different from that in the first 3 months for at least one of the programs, defining any additional symbols you use, or explain why this is not possible.

This question is open to interpretation, so there is no "right" answer, and I gave you credit for how well you explained your answer.

As some of you said, technically, the model as given does not say anything about the form of the relationship of the means over time for each group. The model describes only the means at each time point for each group and does not incorporate a representation of how these means change over continuous time so does not provide a description of the rate of change of mean. Based on this feature alone, some of you said that this is not possible.

However, in Problem 3 of Homework 1, you were asked to express the null hypothesis that the rate of change of mean response is constant for all groups, and it turned out that when the time points are equally spaced as they are here, an expression for this hypothesis *is* possible. Thus, it does not suffice to say simply that expressing the physician's question in this problem is not possible because rate of change is not explicitly represented in the model. You'd need to give a more detailed explanation of why it's not possible.

Some of you used the fact that the time points are equally spaced and expressed the rate of change in the last 2 months for group ℓ, $\ell = 1, 2, 3$, as

$$\frac{\mu_{\ell 6} - \mu_{\ell 4}}{2}$$

and that in the first 3 months as

$$\frac{\mu_{\ell 4} - \mu_{\ell 1}}{3},$$

and thus expressed the null hypothesis as

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

or, equivalently

$$2\mu_{\ell 1} - 5\mu_{\ell 4} + 3\mu_{\ell 6} = 0, \quad \ell = 1, 2, 3.$$

This can be written in terms of $\mathcal{M}$ as

$$\mathcal{M} \begin{pmatrix} 2 \\ 0 \\ 0 \\ -5 \\ 0 \\ 3 \end{pmatrix} = \mathbf{0}.$$

This is an entirely reasonable solution.

Still others of you noted that the visual evidence from the data suggests that the pattern of mean change in the first 3 months looks to be approximately a straight line for each group, so with constant rate of change, and that the pattern in the last 2 months also appears to be approximately a straight line for each group, so again with constant rate of change. You then argued that if the rate of change of mean weight in the first 3 months is the same as that in the last 2 months for all programs, all of the mean profiles must have constant rate of change over the entire study period, and, using the fact that the time points are equally spaced, you expressed this null hypothesis as

$$\mu_{\ell 2} - \mu_{ell1} = \mu_{\ell 3} - \mu_{\ell 2} = \mu_{\ell 4} - \mu_{\ell 3} - \mu_{\ell 5} - \mu_{\ell 4} = \mu_{\ell 6} - \mu_{\ell 5}$$

for $\ell = 1, 2, 3$. This is exactly the same situation as in Problem 3 of Homework 1, so by the same reasoning as in that problem, the hypothesis can be expressed in terms of $\mathcal{M}$ as

$$\mathcal{M} \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \mathbf{0}.$$

This is also a reasonable solution.

[5 points] (c) Based on the information you have, do you feel it is possible to obtain reliable inference on question (i) using model (1)? If so, describe how and present a formal statement of the result. If not, explain why not.

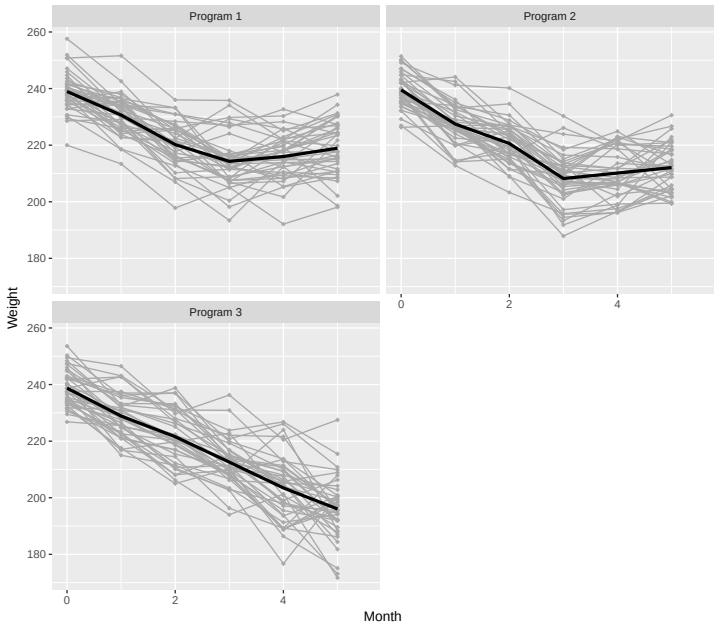
If this were possible, we could address (i) by the test of parallelism, which corresponds to the `month*program` row of the ANOVA table; one would compare the corresponding F ratio to the appropriate critical value. However, there seems to be evidence that the assumptions that must hold for this to be a valid test may be violated. The required assumptions are that the overall covariance matrix is the same for all groups and that this common covariance matrix is at least of Type H.

Although the sample correlation matrices for each group do not seem to be too different and seem all to be consistent with compound symmetry, the test based on Mauchly's criterion of the null hypothesis ("test of sphericity") that the assumed common structure is of Type H strongly rejects this null hypothesis. So if we are willing to believe that there is a common covariance matrix, the evidence suggests strongly that it is not of Type H. It must be that there is enough of a departure from apparent compound symmetry for at least one of the groups that the assumed common patter is not consistent with Type H. If we went ahead with the usual test above under these conditions, it will be too liberal and thus not reliable.

Moreover, inspection of the estimated covariance matrices shows that the variances increase over time in each group. While this is not inconsistent with Type H (look at the definition in (3.16) in the notes), the variances in each group seem to increase differently over time; that for Group 3 increases more dramatically. Perhaps this is a reflection of different overall covariance matrices in each group, which would also violate the required assumptions.

The bottom line is that there is a formal test plus informal evidence that suggest that the usual assumptions required are violated in some way, so that the best answer is "no." Of course, from a pragmatic point of view, the test statistic is pretty large, and the visual evidence strongly suggests that the pattern of change is not the same, so, despite the possible violations of assumptions one probably would be reasonably confident concluding informally that the pattern of change probably isn't the same for all groups.

2. Consider the weight loss study in the previous problem. Here are the data again:



The goals are

(i) To determine if the pattern of change of mean weight is not the same for all programs over the 5 month study period.

(ii) To determine if the rate of change of mean weight in the last 2 months is different from that in the first 3 months for at least one of the programs.

[10 points] (a) Based on all information you have, 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.

All of you recognized that the questions of interest are population-averaged questions, so most of you proposed a population-averaged model directly. Some of you proposed a linear mixed effects model, which is fine, too, as it induces a population-averaged model. Regardless, you recognized that there are two distinct "phases" here in that the weight loss programs were given for 3 months and then stopped, although weight continued to be recorded for 2 more months. Thus, you represented the mean response profile by a linear spline model as in the data analyses in Homeworks 2 and 3.

Here is a direct population-averaged model that addresses the issues. Let Y_{ij} be the weight at the jth month t_j on the ith man, $j = 1, \ldots, 6$, $t_j = 0, 1, 2, 3, 4, 5$, and let $\delta_{i\ell} = 1$ if i was assigned to Program ℓ, $\ell = 1, 2, 3$. Then with x_+ defined as in the course notes, a reasonable model is

$$Y_{ij} = \beta_0 + (\beta_{11}\delta_{i1} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3})t_j + (\beta_{21}\delta_{i1} + \beta_{22}\delta_{i2} + \beta_{23}\delta_{i3})(t_j - 3)_+ + \epsilon_{ij}.$$

An alternative parameterization is

$$Y_{ij} = \beta_0 + (\beta_{11} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3})t_j + (\beta_{21} + \beta_{22}\delta_{i2} + \beta_{23}\delta_{i3})(t_j - 3)_+ + \epsilon_{ij}.$$

Both are fine. In either case, because this is a randomized study, it is reasonable to take the intercept to be common for all groups, although you might have taken it to differ (which would allow a formal test of this and thus of the integrity of the randomization). The among-individual covariates are $\boldsymbol{a}_i = (\delta_{i1}, \delta_{i2}, \delta_{i3})^T$, and there are no within-individual covariates. Letting $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{i6})^T$, assume $E(\epsilon_{ij}|\boldsymbol{a}_i) = 0$ and

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

where $\ell = 1, 2$, or 3 possibly depending on $\boldsymbol{a}_i$. Based on the evidence in Problem 1, we might want to allow a separate covariance matrix for each group. In any case, we assume that $\boldsymbol{\epsilon}_i$ and thus $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})^T$ are independent across i conditional on ba_i. 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 weight under the three programs. Then provide a matrix $\boldsymbol{L}$ such that you can address question (i), to determine if the pattern of change of mean weight is not the same for all programs over the 5 month study period, through an expression of the form $\boldsymbol{L}\boldsymbol{\beta}$.

In either parameterization,

$$\boldsymbol{\beta} = (\beta_0, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23})^T.$$

The $\boldsymbol{L}$ matrix you provided obviously depended on which parameterization you chose. (If you had different intercepts for each group, your $\boldsymbol{\beta}$ included three intercept parameters.)

The pattern of change here is characterized by that in the first phase, through month 3, and then that in the second phase, from month 3 onward. According to the model in either parameterization, it is characterized in the first phase by the group-specific slope; e.g., in the first parameterization, these slopes are β_{11}, β_{12} and β_{13}, while in the second they are

β_{11}, $\beta_{11} + \beta_{12}$ and $\beta_{11} + \beta_{13}$. In the second phase, in the first parameterization the group-specific slopes are $\beta_{11} + \beta_{21}$, $\beta_{12} + \beta_{22}$, and $\beta_{13} + \beta_{23}$, and in the second they are $\beta_{11} + \beta_{21}$, $\beta_{11} + \beta_{12} + \beta_{22}$, and $\beta_{11} + \beta_{13} + \beta_{23}$.

Under the first parameterization, the pattern of change will be the same for all programs is the slopes in each phase are the same for all programs, which will hold if $\beta_{11} = \beta_{12} = \beta_{13}$ and $\beta_{21} = \beta_{22} = \beta_{23}$, in which case

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

For the second parameterization, this holds if $\beta_{12} = \beta_{13} = 0$ and $\beta_{22} = \beta_{23} = 0$, so that

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

The null hypothesis is then $H_0 : \boldsymbol{L}\beta = \boldsymbol{0}$.

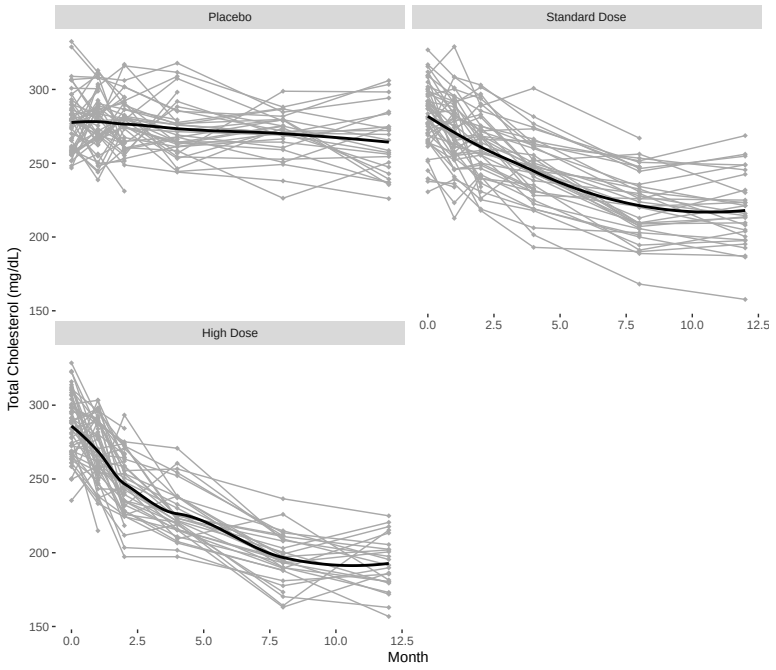
[8 points] (c) In terms of β you defined in (b), provide a matrix $\boldsymbol{L}$ that allows you to address question (ii), to determine if the rate of change of mean weight in the last 2 months is different from that in the first 3 months for at least one of the programs, through an expression of the form $\boldsymbol{L}\beta$.

In either parameterization, it is clear that the slopes in each phase will be the same in all groups if $\beta_{21} = \beta_{22} = \beta_{23} = 0$, so that the null hypothesis is $H_0 : \boldsymbol{L}\beta = \boldsymbol{0}$ where

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

3. The data shown below are from a study of a certain statin medication for use in lowering total cholesterol (mg/dL) in subjects with baseline "high" total cholesterol above 230 mg/dL. 150 subjects were randomized to receive a placebo, the standard dose of the statin, or a high dose of the statin. Total cholesterol levels were to be measured for each participant at baseline (month 0), prior to initiation of assigned treatment, and then at months 1, 2, 4, 8, and 12 thereafter. Also recorded for each subject was an indicator of whether or not the subject had previously taken a statin drug (0 = no, 1 = yes).

   Here are the data, with a loess smooth superimposed on each plot.



s is often the case, many participants dropped out of the study before completion: although all subjects have the baseline cholesterol measurement, only 98 subjects returned for the month 8 measurement, and only 77 returned at 12 months.

The investigators had the following questions:

(i) Is mean baseline total cholesterol level associated with previous statin use?

(ii) Is the mean rate of change of total cholesterol level nonconstant over the 12 month study period for any treatment?

(iii) Is the rate of change in mean total cholesterol level at 6 months associated with treatment received?

[12 points] (a) Can you propose a statistical model in which *all of* questions (i)-(iii) can all 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 at least one of the three questions can be addressed (state which one(s)). Describe (*briefly*) any assumptions you incorporate in the model.

Most all of you noted that questions (i) and (ii) are subject-specific questions (although (i) can also be interpreted from a population-averaged perspective) and (iii) is a population-averaged question. Thus, you said "yes" and proposed a linear mixed effects model, which is subject-specific but induces a population-averaged model, in which all three questions can be addressed. Most of you proposed a model along the following lines:

Let Y_{ij} be the total cholesterol level for subject i, $i = 1, \ldots, m = 150$, at time t_{ij}; because of the possibly missing outcomes, $j = 1, \ldots, n_i \leq n = 6$, where the n intended times are 0, 1, 2, 4, 8, and 12 months. Let $s_i = 1$ if subject i previously used statins and = 0 otherwise, and let $\delta_{i\ell} = 1$ if subject i was randomized to group ℓ and = 0 otherwise, $\ell = 1, 2, 3$, where $\ell = 1$ corresponds to placebo, $\ell = 2$ to low dose, and $\ell = 3$ to high dose.

Then you proposed a hierarchical model, which for most of was a version of the following. The individual model is quadratic

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij},$$

and the population model is of the form

$$\begin{aligned}
\beta_{0i} &= \beta_{00} + \beta_{01} s_i + b_{0i}, \\
\beta_{1i} &= \beta_{11}\delta_{i1} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3} + b_{1i}, \\
\beta_{2i} &= \beta_{21}\delta_{i1} + \beta_{22}\delta_{i2} + \beta_{23}\delta_{i3} + b_{2i}.
\end{aligned}$$

Some of you also allowed the mean of the linear and quadratic coefficients to depend on previous statin use. Most of you took the individual-specific random effects $\boldsymbol{b}_i = (b_{0i}, b_{1i}, b_{2i})^T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$, with $\boldsymbol{D}$ a (3×3) covariance matrix that is the same for all doses (some of you allowed the possibility of a different $\boldsymbol{D}$ for each dose), and $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$ such that $E(\boldsymbol{e}_i | \boldsymbol{a}_i) = \boldsymbol{0}$ and $\text{var}(\boldsymbol{e}_i | \boldsymbol{a}_i) = \boldsymbol{R}_i(\gamma)$ (perhaps $= \sigma^2 \boldsymbol{I}_{n_i}$), where $\boldsymbol{a}_i = (s_i, \delta_{i1}, \delta_{i2}, \delta_{i3})^T$. Some of you allowed a separate mean intercept for each dose as well. The $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ were taken to be independent across i. Some of you parameterized the population model differently, letting placebo be the reference condition for the mean linear and quadratic coefficient.

You needed at the very least to write down a basic model and state basic assumptions about $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ as above to receive full credit.

[7 points] (b) In terms of your model in (a), show how you would address question (i) (is mean baseline total cholesterol level associated with previous statin use?). If you cannot, state why not.

Most all of you correctly proposed testing the null hypothesis that $\beta_{01} = 0$ or the analogous hypothesis for the model as you parameterized it.

[7 points] (c) In terms of your model in (a), show how you would address question (ii) (Is the mean rate of change of total cholesterol level nonconstant over the 12 month study period for any treatment?) If you cannot, state why not.

We talked a lot in class about how to think about "rate of change" in the context of a quadratic model. In the above model, the mean rate of change for treatment group ℓ at any time t is

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

which is of course nonconstant over time. This mean rate of change is constant (i.e., the same for all times t) if $\beta_{2\ell} = 0$. Thus, we can address this question by testing the null hypothesis that

$$\beta_{21} = \beta_{22} = \beta_{23} = 0;$$

if this null hypothesis is rejected, then there is evidence in the data that the mean rate of change is nonconstant for at least one of the groups.

[7 points] (d) In terms of your model in (a), show how you would address question (iii) (Is the rate of change in mean total cholesterol level at 6 months associated with treatment received?). If you cannot, state why not.

In the above model, the rate of change of mean total cholesterol for treatment group ℓ at 6 months is $\beta_{1\ell} + 12\beta_{2\ell}$ (setting $t = 6$ in the expression in (c)). If rate of change at 6 months does not depend on treatment received, then it must be that

$$\beta_{11} + 12\beta_{21} = \beta_{12} + 12\beta_{22} = \beta_{13} + 12\beta_{23}.$$

So to address this question, we want to test this null hypothesis.

Some of you instead specified that we must have $\beta_{11} = \beta_{12} = \beta_{13}$ and $\beta_{21} = \beta_{22} = \beta_{23}$. This would make the rate of change the same for all treatments at all times (so certainly at 6 months). This answer is overkill; all we need is the condition above to hold (at 6 months).

[7 points] (e) Show how you would use your model to estimate the variation in individual-specific rates of change in total cholesterol at the end of the study (month 12) for each treatment, or explain why you cannot do this.

For subject i, the individual-specific rate of change at the end of the study is

$$\beta_{1i} + 24\beta_{2i}.$$

Variation in individual-specific rates of change at 12 months is thus characterized by the variance of this quantity under our model assumptions. From the model, this quantity can be written as

$$\beta_{1\ell} + 24\beta_{2\ell} + b_{1i} + 24b_{2i}$$

if subject i was in treatment group ℓ. Thus, the variance of $\beta_{1i} + 24\beta_{2i}$ is equal to

$$\text{var}(b_{1i} + 24b_{2i}).$$

A few of you left it at this. Others of you wrote

$$\text{var}(b_{1i} + 24b_{2i}) = \text{var}(b_{1i}) + 24^2\text{var}(b_{2i}),$$

forgetting about the covariance between b_{1i} and b_{2i}. In any case, the correct variance

$$\text{var}(b_{1i}) + 24^2\text{var}(b_{2i}) + 2(24)cov(b_{1i}, b_{2i}) = D_{22} + 576D_{33} + 48D_{23},$$

which can be estimated by substituting the estimates of the components of $\boldsymbol{D}$ obtained by maximum likelihood (under the conditions in (f) below). Some of you wrote this correctly in matrix form; defining $\boldsymbol{c} = (0, 1, 24)^T$, you wrote $\boldsymbol{c}^T\boldsymbol{D}\boldsymbol{c}$, which can be estimated by substituting the maximum likelihood estimate of $\boldsymbol{D}$.

[8 points] (f) Under what conditions would you feel comfortable proceeding with a standard analysis using your model? If you would not feel comfortable under any conditions, explain why.

Clearly, the major issue here is the dropout. At the least, we would have to be willing to assume that the dropout is according to a missing at random (MAR) mechanism. This might be the case if, for example, subjects dropped out because their total cholesterol levels at previous visits were not becoming lower on their assigned treatments, and they or their physicians decided they should leave the study and try something else. If we could be reasonably assured that the reasons for dropout are based on information that is available to us in the data, then MAR is reasonable. Under MAR, a valid analysis can still be achieved using maximum likelihood for "large" m as long as we are willing to assume that total cholesterol levels are exactly normally distributed given covariates, our model is exactly correct, and we use model-based standard errors, ideally calculated based on the observed information matrix. Likelihood ratio tests comparing nested models, which could be used to address these questions, would also be valid under these conditions.

I gave full credit if you mentioned the missingness mechanism, maximum likelihood/normality, and something about standard errors/likelihood ratio tests.

[8 points] 4. Suppose that the outcome Y is a measure of growth, and consider the model

$$Y_{ij} = \frac{\beta_{1i}}{1 + \exp\{-(\beta_{3i} + \beta_{2i} t_{ij})\}} + e_{ij}, \quad i = 1, \ldots, m, \;\; j = 1, \ldots, n, \tag{2}$$

$$\beta_{1i} = \beta_1 + b_{1i}, \quad \beta_{2i} = \beta_2 + b_{2i}, \quad \beta_{3i} = \beta_3, \quad \beta_1, \beta_2 \gg 0, \beta_3 > 0,$$

where $\boldsymbol{b}_i = (b_{1i}, b_{2i})^T$ are independent for all i, e_{ij} are independent for all i, j, and $\boldsymbol{b}_i$ and e_{ij} are independent of one another for all i, j, with

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}), \;\; \boldsymbol{D}\,(2 \times 2), \qquad e_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Dick refers to β_1 in (2) as the mean asymptotic growth value among individuals in the population, while Jane refers to β_1 as the asymptotic growth value for the population mean response. Who is correct, Dick or Jane? Explain (*briefly*) your answer.

This was a bit of a trick question. This is a subject-specific model, and β_{1i} is the individual-specific asymptotic growth value, with mean β_1. Thus, β_1 does have the interpretation as the mean asymptotic growth value among individuals in the population, so it is immediate that Dick is correct.

Jane is trying to give β_1 a population-averaged interpretation – under this perspective, β_1 should be the asymptotic growth for the population mean growth profile. From the model, the population mean growth profile is

$$E\left[\frac{\beta_{1i}}{1 + \exp\{-(\beta_3 + \beta_2 t_{ij} + b_{2i} t_{ij})\}}\right]$$
$$= E\left[\frac{\beta_1}{1 + \exp\{-(\beta_3 + \beta_2 t_{ij} + b_{2i} t_{ij})\}}\right] + E\left[\frac{b_{1i}}{1 + \exp\{-(\beta_3 + \beta_2 t_{ij} + b_{2i} t_{ij})\}}\right].$$

From Problem 2 of Homework 1, you know that the first term in this expression is approximately a logistic function with asymptotic growth value β_1. Several of you concluded that the second expectation in this expression, which is of the form

$$E\left[\frac{b_{1i}}{1 + \exp\{-(A_{ij} + b_{2i} t_{ij})\}}\right] \tag{3}$$

for A_{ij} a constant, is equal to zero because $E(b_{1i}) = 0$, and thus concluded that Jane is also correct on that basis (several of you said that the model is linear in b_{1i}). However, the expression inside this expectation also depends on b_{2i}, and b_{1i} and b_{2i} are correlated, so it is not clear that this expectation is equal to zero (it probably isn't), and thus this answer is not necessarily correct. I did not expect you to work out this expectation, of course! It is very likely that the expectation of (3) converges to zero as $t_{ij} \to \infty$, given the nature of the dependence of the integrand on t_{ij}, in which case β_1 would be the asymptotic growth value for the population mean profile, making Jane correct.

I gave full credit for any answer as long as you explained it.