1. Suppose that we have a situation in which experimental units have been randomized into $q = 3$ groups and each unit is observed at the same $n = 4$ times. As in the notation in the notes, let $\mu_{ij}$ be the mean for the $i$th group at the $j$th time.

(a) In terms of the $\mu_{ij}$, write down the M matrix for this problem.

(b) Suppose $y_i$ is the data vector for the $i$th experimental unit. If we assume the model given in equation (4.11), write down the vector $a_i$ appropriate for an experimental unit in group 2.

(c) Suppose $\mu_{ij} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij}$. If we express the model in the alternative "regression form" given on page 119 of the notes, $y_i = X_i\beta + \epsilon_i$, write down the vector appropriate vector $\beta$ and the "design matrix" $X_i$ for the $i$th experimental unit in the 3rd group.

(d) Write down the appropriate matrix $C$ that would be used in this situation to characterize a set of contrasts focusing on difference among the $q = 3$ groups.

(e) Write down the matrices $C$ and $U$ that could be used to represent the null hypothesis that the mean profiles are parallel across the $q = 3$ groups in the form $H_0 : CMU = 0$.

2. Estimation of $\sigma_b^2$ and $\sigma_e^2$. For a general problem with $q$ groups and $n$ time points, with $m$ total experimental units, suppose we construct the usual univariate repeated measures analysis of variance table given on the top of page 130 of the notes. Suppose that the value of mean square corresponding to among-unit error, $MS_{EU}$, has numerical value $a$, and the mean square corresponding to within-unit error, $MS_E$, has a numerical value $c$.

(a) Consider the table of expected mean squares on the top of page 131. How would you estimate $\sigma_b^2$ and $\sigma_e^2$ given the values $a$ and $c$? Answer this question by providing expressions for the estimates, $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$, say, in terms of the values $a$ and $c$.

(b) Suppose that $c = 100$, $a = 25$, and $n = 4$. Based on (a), provide a numerical estimate of the matrix $\Sigma$. 

3. In the file cd4.dat found on the class home page, you will find longitudinal data on m = 50 subjects infected with HIV virus. Each subject was randomly assigned to one of 2 groups: Subjects 127 (group 1) received one drug (monotherapy); Subjects 2850 (group 2) received this drug in combination with two additional drugs (triple therapy). At time 0 corresponding to initiation of treatment and at weeks 8, 16, 24, and 32 thereafter, a measurement of CD4 lymphocyte count was obtained from each subject. (CD4 lymphocyte count is a standard measure of the state of the immune system and is routinely collected in studies of HIV-infected individuals; the higher the count, the better.) Note that the data set is actually in the form needed for using the repeated statement in proc glm. To get it in the form with one response on each line, use

```plaintext
data cd41;
  infile ...\cd4.dat; /* add the location */
  input subject week1 week2 week3 week4 week5 group;
run;

data cd42; set cd41;
  array e(5) week1 week2 week3 week4 week5;
  do week = 1 to 5;
    logcd4 = e(week);
    output;
  end;
  drop week1 week2 week3 week4 week5;
run;
```

The investigators were interested in whether subjects in the two treatment groups exhibited different patterns of logarithm of CD4 count over time. The data set has 7 columns:

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Subject ID number</td>
</tr>
<tr>
<td>2-6</td>
<td>Log CD4 counts for weeks 0, 8, 16, 24, 32</td>
</tr>
<tr>
<td>7</td>
<td>Group indicator (1=monotherapy, 3=triple therapy)</td>
</tr>
</tbody>
</table>
(a) Write a SAS program to obtain the following analyses:

(i) Read in the data in the form they appear in the data set

(ii) Find the means for each group at each day and plot them for each group on the same graph

(iii) Conduct a univariate repeated measures ANOVA using PROC GLM via the split-plot specification of the model; this will require you to create a new data set with one data record per subject/week. Use a random statement to get SAS to print out the table of expected mean squares.

(iv) Conduct a univariate repeated measures ANOVA using PROC GLM using the repeated statement. Have your program print out the test for sphericity.

(v) Use PROC GLM and appropriate repeated statements to obtain specialized within-unit analyses using the polynomial, profile, and Helmert transformations.

(vi) Construct a new variable that is the average of the Log CD4 counts for all 5 weeks. Use PROC TTEST to test the mean group difference for this average variable. On the basis of the output, answer (b)(i) below.

On the basis of the output, answer (b)(i) below.

(b) Informally, from the plot of the group/day means (ii), do you think that the visual evidence supports the contention that the pattern of log CD4 count over time is similar? Explain your answer.

(c) From the output of (iii), write down the numerical value of the $F$ ratio suitable for testing whether change in mean log CD4 count is different for the two drug regimens. At level significance $\alpha = 0.05$, is there sufficient evidence to reject this hypothesis? Provide justification for your answer.

(d) From the output of (iii) and using your answer to problem 2 above, provide estimates of $\sigma^2_b$ and $\sigma^2_e$ under the assumption of the compound symmetry model, and write down the estimate for $\Sigma$. Provide an estimate of the correlation among two measurements on the same subject.
(e) Is there sufficient evidence in these data to conclude that the assumption of compound symmetry is inappropriate? Justify your answer by citing the appropriate results from the output of (iv). Is it necessary to use the adjusted $F$ test for parallelism here?

(f) Does it make sense to consider the test of whether the mean response differs between the two groups averaged across time? Why or why not?

(g) Regardless of your answer to (f), write down the numerical value of the $F$ ratio and $p$-value that would be appropriate for addressing this issue. From the output of PROC TTEST in (vi), give the pooled t-value and associated $p$-value. Comment on the similarity of the two $p$-values.

(h) From the output for (v), give the numerical value of the $F$ test statistic that would be appropriate for testing whether the two groups differ in terms of linear trend over time. Is there sufficient evidence to conclude at level $\alpha = 0.05$ that the mean log CD4 count patterns for the 2 groups differ in this way? From the plot in (ii), does the result of this test seem to agree with the visual evidence? Explain your answer.

(i) From the output for (v), give the numerical value of the $F$ test statistic that would be appropriate for testing whether, on average across the two groups, there is evidence of a quadratic component to the trend over time. Is there sufficient evidence to conclude at level $\alpha = 0.05$ that such a feature exists?

(j) The investigators suspected that the groups might show the most striking differences toward the end of the study; thus, they were interested in whether the way in which mean responses at weeks 24 and 32 differ is in fact different for the 2 treatments. From the output for (v), give the numerical value of the $F$ test statistic that would be appropriate for testing this issue. Is there sufficient evidence to conclude at level $\alpha = 0.05$ that the mean log CD4 count difference between weeks 24 and 32 is different for the two groups? From the plot in (ii), does the result of this test seem to agree with the visual evidence? Explain your answer.

4. (a) Write a SAS program to obtain the following analyses for the CD4 count data in problem 3:
(i) Read in the data.

(ii) Use PROC CORR to obtain estimates of the covariance matrix and associated correlation matrix for each group separately.

(iii) Use PROC GLM with the MANOVA statement to obtain the various tests of equality of the group mean vectors. Print out the SS&CP matrices $Q_H$ and $Q_E$ corresponding to this hypothesis.

(iv) Call PROC GLM again with the repeated statement to obtain the profile analysis.

(b) From (ii), write down the estimated covariance matrix and its associated correlation matrix for each group. From informal inspection of these estimates, does the assumption of compound symmetry seem reasonable for each group? Does the assumption that the pattern of variation is similar within each group seem to be supported by the estimates? Explain your answers.

(c) In the output, write down an estimate of the correlation matrix of a data vector under the assumption that this matrix is the same across groups. Does this estimate support the assumption of compound symmetry?

(d) From (iii), do the various test statistics suitable for testing the null hypothesis of whether the mean vectors differ across time, i.e. $H_0 : \mu_1 = \mu_2$, yield the same result? Why or why not? Based on the results of these tests, is there sufficient evidence assuming no particular structure for the matrix $\Sigma$ to conclude at level $\alpha = 0.05$ that the means do differ somehow? Give the value of the test statistic(s) you use to justify your answer and explain why they lead you to make this conclusion.

(e) From (iv), give the value of the test statistic(s) on which the multivariate test of parallelism is based. At level $\alpha = 0.05$, is there sufficient evidence assuming no particular structure for $\Sigma$ to conclude that the mean profiles are not parallel? Justify your answer based on the output.