BIOS6643 HW3 Due Friday, October 4, 2018, 5pm

(1) Consider a basic science experiment conducted where cell counts are measured at 4 time points for samples taken from individual subjects or animals. A linear mixed model will be fit for the data (perhaps after log transformation), and fixed effects will be included for time, and possibly treatment group as well as their interaction. (To answer this question we do not need to know the specific form of **Xβ**.) Determine the structure for **V**$_i$ if a random intercept for subjects will be included, plus an AR(1) structure for the error covariance matrix (**R**$_i$). What does the combination of non-simple **R** and **G** allow you to do in modeling covariances that using only one cannot do? Discuss in a few sentences.

Solution: Recall that $Var(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i^{\,t} + \mathbf{R}_i$. Here, **Z**$_i$ is just a column of 1's since there is only a random intercept, so $\mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i^{\,t} = \sigma_b^2\mathbf{J}$. Also, $\mathbf{R}_i = \sigma_\varepsilon^2\begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{pmatrix}$ based on the AR(1), so we have

$$Var(\mathbf{Y}_i) = \sigma_b^2\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \sigma_\varepsilon^2\begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{pmatrix} = \begin{pmatrix} \sigma_b^2+\sigma_\varepsilon^2 & \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2\phi^2 & \sigma_b^2+\sigma_\varepsilon^2\phi^3 \\ \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2 & \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2\phi^2 \\ \sigma_b^2+\sigma_\varepsilon^2\phi^2 & \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2 & \sigma_b^2+\sigma_\varepsilon^2\phi \\ \sigma_b^2+\sigma_\varepsilon^2\phi^3 & \sigma_b^2+\sigma_\varepsilon^2\phi^2 & \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2 \end{pmatrix}$$

The advantage of this structure is that you get some decay between covariances as the gap between time points is increased, but it does not necessarily decay to 0. This is helpful for correlated longitudinal data where some correlation between responses is expected regardless of gap between responses, but it is not as simple as the CS structure, either. Of course, we might be able to find covariates that explain the subject heterogeneity, which might then eliminate the need for the random intercept component.

(2) One model we used for the Mt. Kilimanjaro data included random effects for subject, up to the quadratic term (plus covariances between random effects), along with a simple **R** structure. (We did find at least one model with a better AIC, but let's focus on this one for now.) We talked about how including multiple random effects can induce a covariance structure that is time sensitive (or in this case, altitude sensitive). Show this by considering a simple data set and model. In particular, let times be *t*=0, 1, 2, and consider a model that includes a random intercept and slope for time by subject, plus covariance between them (i.e., UN structure in **G**). Show that it is possible to obtain Cov(Y_{i1},Y_{i2}) > Cov(Y_{i1},Y_{i3}) < Cov(Y_{i2},Y_{i3}), i.e., decaying covariance as distance between time points is increased. For what covariance parameter values will these hold?

Solution: $\mathbf{Z}_i = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$, so

$$\mathbf{V}_i = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} + \sigma_\varepsilon^2\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} \sigma_0^2+\sigma_\varepsilon^2 & \sigma_0^2+\sigma_{01} & \sigma_0^2+2\sigma_{01} \\ \sigma_0^2+\sigma_{01} & \sigma_0^2+2\sigma_{01}+\sigma_1^2+\sigma_\varepsilon^2 & \sigma_0^2+3\sigma_{01}+2\sigma_1^2 \\ \sigma_0^2+2\sigma_{01} & \sigma_0^2+3\sigma_{01}+2\sigma_1^2 & \sigma_0^2+4\sigma_{01}+4\sigma_1^2+\sigma_\varepsilon^2 \end{pmatrix}$$

To get the 'decay' pattern we need $0 > \sigma_{01} > -2\sigma_1^2$. Note that since variances are changing over time, we could also consider finding constraints on the correlations, which might be even more meaningful. However, this is an algebraic nightmare and would be aided by some algebra software.

(3) Fit the Beta Carotene data using a continuous model for time, including group and group*time in the model. (For a description of the data, see the file in the HW folder.) Determine the degree of polynomials for time that is important and sufficient for the model. For covariance structure, define the UN structure for **R**.

a. Write your final model, fit it and compare it to the model that used group, time and group*time as class variables. Which would you go with in a final report? Explain. NOTE: in comparing model AIC's use method=ML for a more apples-to-apples comparison, particularly when changes are being made to the fixed effects.

We should include relevant polynomial terms for both time and group*time. For example, in a cubic model we would include time, time*time and time*time*time, and also group*time, group*time*time and group*time*time*time. You can do some model reduction if you like, but if so, you want to eliminate higher order terms first. Here, I will keep all terms for the relevant polynomial model (for both time and group*time).

Since there are 5 time points we know that we don't need to look past the quartic model, since that model is equivalent to the model using group and time as class variables. This is because in those cases you 'saturate' the model, making estimates for group and time combinations as flexible as possible. Remember that you can draw a straight line through 2 points, a quadratic through 3 points, a cubic through 4, and a quartic through 5.

The AIC's for the linear, quadratic, cubic and quartic models are 1252.0, 1246.1, 1243.1 and 1245.3, respectively. (The model using group and time as class variables also yields 1245.3 because of what was discussed in the last paragraph.) This suggests the cubic model is the best, based on AIC. Given the AIC's are pretty close, you could probably go with one or the other depending on other criteria. For example, you argue that the class predictor approach is reasonable here, but if you want to interpolate between times, you might consider the cubic model.

<u>The LTFR cubic model is</u>

$$Y_{hij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 + \alpha_h + \gamma_{1h} x_{ij} + \gamma_{2h} x_{ij}^2 + \gamma_{3h} x_{ij}^3 + \varepsilon_{hij},$$

where *h*, *i* and *j* denote group, subject and time, respectively. If we compile errors for a subject into vector ε_{hi}, we have ε_{hi}~N(0,R$_i$), where R$_i$ has the UN structure. We could further trim off the *i* subscript on *x* if all subjects have the same time points. If they actually had slightly different time points, keep it, though.

b. Write an estimate or contrast statement for your continuous model based on what you think is interesting. The custom estimate and/or test could involve a subset of the data (e.g., comparing 2 specific groups), or the whole data.

Below is the SAS code for the curve comparison for the 2 BASF groups. Note that this looks a little different than the curve comparison in the notes (Mt. K application) since I am treating prepar as a class variable. When I say 'curve comparison', I mean anything that will differentiate the functions, including intercept differences. Thus, this is a very general test. If you have any questions on this, let me know. Note that I kept the ML method here that was used to compare models; it could be changed back to REML for inference purposes, which would provide a slightly more accurate (less biased) test result.

```
*LTFR model;
proc mixed data=univar method=ml;
class id prepar;
model y= prepar time time*time time*time*time prepar*time prepar*time*time
  prepar*time*time*time / solution;
repeated / subject=id(prepar) type=un;
contrast 'BASF curve comparison' prepar 0 0 1 -1, prepar*time 0 0 1 -1,
prepar*time*time 0 0 1 -1, prepar*time*time*time 0 0 1 -1; run;
```

```
                              Contrasts

Label                        Num DF   Den DF   F Value   Pr > F

BASF curve comparison           4       19       4.56    0.0095
```

(4) Consider a study where children are sampled from schools, and then measured over time. We will include a random intercept for schools and for subjects within schools (but simple **R**). Determine $\mathbf{V}_h$, the covariance matrix for school *h*, if there are 3 children sampled from this school, where the first two kids have 3 measures and the last has 2. You might find it helpful start by writing the model for outcome Y_{hij} and determining the design matrix for the random effects. (You can just write something generic for the fixed-effect part of the model.) <u>For thought, not to turn in</u>: how would $\mathbf{V}_h$ change if we had more measures for subjects and employed the AR(1) structure for $\mathbf{R}_{i(h)}$ (the error covariance structure for subject *i* within school *h*)?

$$\mathbf{V}_h = \mathbf{Z}_h\mathbf{G}_h\mathbf{Z}_h^t + \mathbf{R}_h\,,\ \mathbf{R}_h = \sigma_\varepsilon^2\mathbf{I}\ ,;\quad \mathbf{Z}_h = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \mathbf{G}_h = \begin{pmatrix} \sigma_F^2 & 0 & 0 & 0 \\ 0 & \sigma_S^2 & 0 & 0 \\ 0 & 0 & \sigma_S^2 & 0 \\ 0 & 0 & 0 & \sigma_S^2 \end{pmatrix}.$$

Combining these, we get

$$\mathbf{V}_h = \begin{pmatrix}
\sigma_F^2+\sigma_S^2+\sigma_\varepsilon^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 \\
\sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2+\sigma_\varepsilon^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 \\
\sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2+\sigma_\varepsilon^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 \\
\sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2+\sigma_S^2+\sigma_\varepsilon^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2 & \sigma_F^2 \\
\sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2+\sigma_\varepsilon^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2 & \sigma_F^2 \\
\sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2+\sigma_\varepsilon^2 & \sigma_F^2 & \sigma_F^2 \\
\sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2+\sigma_S^2+\sigma_\varepsilon^2 & \sigma_F^2+\sigma_S^2 \\
\sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2+\sigma_S^2 & \sigma_F^2+\sigma_S^2+\sigma_\varepsilon^2
\end{pmatrix}$$