BIOS6643 **Midterm** **2016**

(1) The fitted $\mathbf{R}_i$ matrix for one subject in the Beta Carotene data has the following form when using the UN structure ($\mathbf{R}_i$ matrix on left, $\mathbf{Rcorr}_i$ on right). This is the same as $\mathbf{V}_i$ as there are no random effects in this particular model. Recall that the 5 time points were at 0, 6, 8, 10 and 12 weeks. (We just used the second BL value in this analysis.)

a. Based on observing the numbers in the matrix, argue why you think the AR(1) either would work or would not work for these data; 2 to 3 sentences is sufficient.

Estimated R Matrix for Id(Prepar) 71 1

Col1	Col2	Col3	Col4	Col5
2825	4114	5334	3939	4304
4114	14162	13215	11600	12287
5334	13215	14514	11697	12450
3939	11600	11697	11396	11313
4304	12287	12450	11313	12552

Estimated R Correlation Matrix for Id(Prepar) 71 1

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.6505	0.8330	0.6943	0.7229
2	0.6505	1.0000	0.9218	0.9131	0.9216
3	0.8330	0.9218	1.0000	0.9095	0.9224
4	0.6943	0.9131	0.9095	1.0000	0.9459
5	0.7229	0.9216	0.9224	0.9459	1.0000

For the AR(1) structure, you would force correlation to decay (given ϕ is positive) as time between measurements is increased. We do not see that pattern here. Also, in the AR(1) we assume that correlation between measurements that are equally spaced will be the same, e.g., those indicated by the red line. That is clearly not happening here; those spaced one time point apart go from 0.65 up to about 0.95. (10 points.)

b. Say that we model the data in a different way; we include the baseline value as a covariate rather than as an outcome. How would you expect this to affect the selection of a correlation structure ($\mathbf{V}_i$) for the responses? Again, 2 to 3 sentences is sufficient.

Note that the baseline measure occurs much earlier in time and before treatments are applied, which can contribute to differences in $\mathbf{V}_i$ we see when it is included as an outcome. You need to consider what the subset of $\mathbf{V}_i$ would look like that eliminates the first row and column above. It could be that the covariance matrix will be affected by the fact that we're fitting the model with a new predictor. However I would expect that by examining the subset of $\mathbf{V}_i$ above that eliminates the first row and column, we can get an idea of what might happen with the new model. Note that we now have 4 times during the treatment phase that are equally spaced. It is likely that a simpler structure (CS, Toeplitz, AR(1)) would be adequate, or at least closer to the UN in terms of goodness-of-fit, based on what we see above. If we tried CS, we see that the correlations in the off diagonal are in the 0.91 to 0.95 range.

An alternative to letting the BL value be a covariate would be to try a spatial structure that has fewer parameters than the UN, and accounts for the unequal spacing.

(10 points.)

O.k. so after answering the question, here is what we get when rerunning the suggested way (this is more just to see if our intuitions are correct):

Estimated R Matrix for Id 71

Row	Col1	Col2	Col3	Col4
1	9296.40	5875.56	7088.60	7035.69
2	5875.56	4700.47	4712.24	4724.46
3	7088.60	4712.24	7239.18	6415.87
4	7035.69	4724.46	6415.87	6914.34

Estimated R Correlation Matrix for Id 71

Row	Col1	Col2	Col3	Col4
1	1.0000	0.8888	0.8641	0.8776
2	0.8888	1.0000	0.8078	0.8287
3	0.8641	0.8078	1.0000	0.9069
4	0.8776	0.8287	0.9069	1.0000

The AIC for the UN structure is now 840.8. For TOEPLITZ, we get AIC=838.7, for AR(1), AIC=845.6; CS yields 835.0. So the findings match what we expected, more or less. You could try to add a random intercept and slope to see what happens, as an alternative to modeling R.

c. Write an ESTIMATE statement (to be added to PROC MIXED code) to compare the linear trend for Preparation 1 versus linear trend for Preparation 3 (group, time as class variables and interaction between them included); indicate which type of model you are using (e.g., means, 2-way effects). Note that there are 4 total preparations and you now just have 4 times.

ESTIMATE ‘lin trend, prep 1 v 3’ group*time -3 -1 1 3 0 0 0 0 -3 -1 1 3 0 0 0 0;

This is actually the same for the means and 2-way effects models. There are other coefficients that would work to complete the test, but the formulation above allows it to be orthogonal to other polynomial trends (since they are centered on 0). (10 points.)

(2) Say we find 4 different schools to participate in ‘health day’, and within each school we find 5 children to take blood pressure measurements on.

a. Are schools and children nested or crossed factors? **Nested (5 points.)**

b. Suppose that we build a model for blood pressure that includes age and gender as covariates. Write a complete observation-level statistical model for these data if we take just 1 measurement per subject, and either add terms or specify structures to address potential correlation between responses in your model.

Let h denote school, i subject, k gender. Here is my model:

$$Y_{hik} = \beta_0 + \beta_1 x_{1i} + \alpha_k + b_h + \varepsilon_{hik}$$

where $b_h \sim N(0, \sigma_S^2)$**, independent of** $\varepsilon_{hik} \sim N(0, \sigma_\varepsilon^2)$**;** x_{1i} **is age (assuming *i* is a unique study-wide subject index and we’re using standard LMM methods), alpha parameters are for gender. This is a less-than-full-rank model. (10 points.)**

c. Repeat the previous part if we take 2 measurements per subject.

So I gave you some freedom here to decide whether to include another factor for time or not. If you don't include it, one model might be

$$Y_{hijk} = \beta_0 + \beta_1 x_{1ij} + \alpha_k + b_h + b_{i(h)} + \varepsilon_{hijk}$$

where h is for school, i for subject, j for repeated measure, and k for gender, $b_{i(h)} \sim N(0, \sigma_P^2)$, independent of other random terms, which are the same as defined in part b. Recall that adding a random intercept induces correlation on the next smaller unit. So even if we let the errors be *iid*, we have accounted for correlation on the repeated measures in some way. Given that there are only 2 measures per subject, this might be sufficient. But we could try adding an UN structure for subject, in which case we would then drop the random intercept for subject (within school). (10 points.)

d. Determine the covariance between time measurements within one subject for this model. (You don't need to compute the entire covariance matrix unless you want to.)

For the model above, it would be

$$Cov(Y_{hijk}, Y_{hij'k}) = Cov(b_h + b_{i(h)} + \varepsilon_{hijk}, b_h + b_{i(h)} + \varepsilon_{hij'k}) = Var(b_h) + Var(b_{i(h)}) = \sigma_S^2 + \sigma_P^2$$

(10 points.)

Short answer questions: Select 2 of the following questions to answer; 2 to 3 sentences and a simple illustration (if applicable) is sufficient. FIRST, CIRCLE THE TWO QUESTIONS THAT YOU WILL ANSWER, AND THEN WRITE YOUR ANSWERS IN THE EMPTY SPACE BELOW. IF YOU NEED MORE SPACE, USE AN ADDITIONAL BLANK PAGE RATHER THAN CONTINUING ON THE BACK. **(10 points for each)**

(3) We fit a model for longitudinal data and use the UN structure for the R matrix. We then add a random intercept to this model but one of 3 things happens: (i) the final Hessian matrix was not positive definite, (ii) model did not converge, (iii) variance of the random intercept is estimated to be 0. (It might be any of the 3, it just depends on the data.) But then if we then change the structure for R to simple ($\sigma^2\mathbf{I}$) and keep the random intercept, the convergence criteria are met and the variance of the random intercept is sizeable. Explain what is happening. If you use an example, a simple matrix will do (e.g., 2 repeated measures per subject).

As we discussed in class, if you put an UN structure on subjects, then including a random intercept in addition to that will lead to identifiability issues, which can easily be shown with a 2x2 matrix. If you try to fit such an overparameterized model, in some cases one covariance parameter will be estimated to be 0, and in other cases you will have more serious non-convergence or error messages. Even in case (iii), you should remove the intercept to help make sure the model fits well.

(4) A friend of yours fits data using a linear model. You then remind them that they actually have longitudinal data, and that they should use some kind of model that takes correlation between repeated measures into account. They then say, "Yeah, I tried that but the estimates did not change much, so I'm going to stay with the simpler model." At first you think, oh, o.k. But then you remember, "Ah, but…" Finish this statement (i.e., explain why you the longitudinal model is expected to be better despite the fact that the estimates themselves may not be much different between the approaches).

We discussed in class how often changing covariance structure (which could include going from 'simple' as is there were no correlated data to something more complex) often does not change estimates very much, but that SE's can be more impacted. Tests can also be affected by (D)DF selection, which depend on how you deal with the correlated data in the model. These things said, for some applications I have also seen estimates themselves change based on how correlated data is handled in the model; recall that in ML estimation, fixed-effect and covariance parameters are estimated simultaneously. Afterthought: for several of the datasets we've already analyzed, when I simply drop the correlation structure and assume independent measures, the SE's of the Beta estimates generally increase and the DF's also increase due to the default selection that assumes independent measures. However, in general, when you misspecify the covariance structure, the SE's can either go up or down. The important thing is getting a decent covariance structure so that resulting SE's are not either over or underestimated too much.

(5) In class we observed when we fit the Mt. Kilimanjaro data using PROC MIXED that the variance for the random intercept was 0 (and hence essentially that parameter was dropped from the model). However, we still obtained nonzero estimates using the EB method. These seem to conflict to each other. Describe why this may happen when fitting a mixed model and why it doesn't necessarily indicate a 'bad' fit.

It's mainly because the estimation of parameters and random effects is done in 2 different steps, one uses the marginal model, and one uses the conditional model (given random effects). Estimation of parameters in the model is accomplished using ML or REML (usually), and it uses the marginal model, while random effects are estimated using empirical Bayes methods, which employs the conditional distribution. Recall that the EB estimates are a weighted average of a subject's own data and the population average estimate.