Stat 771 Homework 4, due Monday, April 18

We will consider data from an AIDS trial comparing two drugs, ddI and ddC (Abrams et al., 1995). This was a multicenter, randomized, open label trial to compare ddI to ddC in HIV-infected patients who were intolerant to or had failed ZDV therapy. Participants were men and women, aged 15–67, who had either advanced to AIDS or had a CD4 lymphocyte count ≤ 300 (this is low). The primary objective is to find out which therapy is better, ddI or ddC, at increasing lymphocyte counts; CD4 T-cell counts are widely-used biomarkers measuring the progression of HIV infection and AIDS (low cells counts are the hallmark of AIDS). There are $m = 467$ patients that could *potentially* be observed at 5 time points: 0, 2, 6, 12, and 18 months after therapy started. At 0 months, no there are no differences in the populations, and we will include the CD4 count at 0 months as a `base`line measurement in the `model` statement. Instead of working with the CD4 counts directly, we will take the square root and work with $Y_{ij} = \sqrt{CD4_{ij}}$. The square root is a common variance stabilizing transformation for counts, often used with Poisson data.

This is a fairly large, messy data set with missing observations, and observations taken at unequally spaced times. The overarching goal is to find out whether there is a treatment effect (`randgrp` $g_i = 1$ for ddI $g_i = 2$ for ddC), while controlling for baseline CD4 counts (`base` b_i), the `stratum` effect ($s_i = 1$ is ZDV failure and $s_i = 2$ is ZDV intolerant), and clinic (`unit` $u_i = 2, \ldots, 18$) effects. We want to build a model that allows for flexibility in covariance structure, but does not totally overfit the data. A random model coefficient with linear trends might work well here.

1. Obtain profile plots of CD4 count stratified by `randgrp` and `stratum` (four plots total). Given the rather limited amount of information, are linear subject-specific profiles plausible here? Does there seem to be a `randgrp` difference based on the plots (within each level of `stratum`)? You might superimpose a fitted *line* into each of the four plots; you can do this with the `loess` option and picking `smooth=` to be a number that gives a line (`smooth=1`) worked for me).

2. Obtain the OLS intercept/slope estimates for each subject. Plot the intercepts versus each of `randgrp`, `stratum`, `base`, and `unit`. Do the intercepts seem to change with any of these predictors? Repeat for the slopes.

   We are plotting intercepts/slopes versus main effects only. If these plots look like random noise, this does not preclude *interactions* among the predictors/adjusters/confounders affecting the intercept and/or slopes. Below you will fit the model with all interactions and work backwards.

3. Now plot the intercept/slope *pairs* versus the four levels of `randgrp` and `stratum`. Is there any obvious differences across scatterplots? Which of `randgrp` and `stratum` (maybe one, maybe both, maybe neither) affect the *shape* of the intercept/slope distribution? It is these categorial covariances that we can include in a `group=` to model distinct $\mathbf{D}$ matrices.

4. Fit a random coefficient model with all possible covariate by covariate and time interactions, and separate unstructured $\mathbf{D}_{s_i}$ matrices for each level of stratum. e.g.

```
proc mixed data=cd4_2 method=ml;
 class randgrp stratum unit id;
 model cd4=randgrp stratum base randgrp*stratum randgrp*base stratum*base time
       time*randgrp time*stratum time*base time*randgrp*stratum time*randgrp*base time*stratum*base unit / s;
 random intercept time / subject=id type=un group=stratum;
run;
```

   Work your way backwards *hierarchically* eliminating insignificant *higher order* effects based on Type 3 tests until all higher order terms are significant at the 10% level. That is, perform backwards elimination by hand from a full model.

5. Obtain *conditional* studentized residuals from your final model and plot then versus `randgrp`, `stratum`, `base`, and `unit`. Does the residual variability seem to change with any of these predictors?

6. Based on your final model from parts 4 and 5, refit the model with `repeated / local=exp(stratum base);` This fits the model with residual variance depending on `stratum` and `base`, i.e. $var(\mathbf{e}_i) = \exp(\tau_0 + \tau_1 s_i + \tau_2 b_i)\mathbf{I}_{n_i}$. Are these affects significant (put `covtest` into the `proc mixed` statement)? That is, do we reject $H_0 : \tau_1 = 0$ and $\tau_2 = 0$?

7. Obtain default residual plots using ODS graphics and putting `/ residual` in the model statement. Do modeling assumptions appear to be okay here? Elaborate.

8. Carefully describe the `randgrp` effect, adjusting for the remaining variables. In particular, if `randgrp` has an interaction with any other effects, carefully describe the nature of the relationship between `randgrp` and those effects. Otherwise, with no interaction, your job is considerably easier!

9. We have including the `unit` effect as fixed. Instead, now allow the unit effects to be random by including the additional statement `random unit / subject=id;` in the model and taking `unit` out of the `model` statement. Does the AIC increase or decrease?

10. Extra credit. Since the observation times t_0, t_1, t_2, t_3, t_4 do not change from person to person, we can fit a model from Chapter 8 to these data easily. Fit the same mean model but remove the `random` statement(s) and add a repeated statement that allows for compound symmetric (homogeneous and heterogeneous) and unstructured covariance matrices. How does the AIC change? Be careful, count the mean parameters $\boldsymbol{\beta}$ and variance components $\boldsymbol{\omega}$ by hand in each case.