BIOS6643 Homework 6 Solutions 2016

Not to turn in:

A. Practice to model splines. If you did it, please send in your graphs and equations and I can review it.

To turn in:

(1) Patients with Chronic Beryllium Disease have ongoing visits at NJH to monitor their health. Once measure taken during their visits is *AADO₂R* (alveolar-arterial oxygen tension difference at rest; the lower the value, the better the health; see course notes for more detail). The goal is to estimate the effect of the disease on *AADO₂R* over time, after accounting for variables known to be related to it (age, gender, height). One modeling challenge is that subjects come in on different days, have different times between visits, and don't have the same number of visits.

a. Consider the model for *AADO₂R* as a function of time since first exposure (*ntep*, in years), age (*ageep*, in years), *height* (inches) and *gender*; the variables *ntep*, *ageep* and *height* are time-varying; a random intercept is included for subjects. Write out a statistical model.

Solution:

$Y_{ij} = \beta_0 + \beta_1 ntep_{ij} + \beta_2 age_{ij} + \beta_3 height_{ij} + \beta_4 female_i + b_i + \varepsilon_{ij}$ (full rank)

where *i* denotes subject, *j* denotes time; Y=*AADO₂R*; female = 1 for females, 0 for males; $b_i \sim \text{N}(0, \sigma_b^2)$, $\boldsymbol{\varepsilon}_i \sim \text{N}(0, \sigma_\varepsilon^2 \mathbf{I})$. You can also use a variable such as θ_k , k=1,2 (for F, M) in place of $\beta_4 female_i$, inducing a less-than-full-rank model. This is the model SAS fits, which is

$Y_{ijk} = \beta_0 + \beta_1 ntep_{ij} + \beta_2 age_{ij} + \beta_3 height_{ij} + \theta_k + b_i + \varepsilon_{kij}$ (less than full rank)

Be careful not to mix different types of notation. If you use the 'regression' approach (i.e., full-rank model), you need to use a dummy (0/1) variable; if you specify a variable that has values such as 1 or 2, M or F, you need to demonstrate it is a class variable with 2 levels.

b. Using software, fit the model. (Note: *ageep*, *height* and *gender* are 'a priori' covariates. They stay in the model regardless of their significance.)

```
proc mixed data=long.Be_study;
class sex id;
model aado2r=ntep height sex ageep / solution;
random intercept / subject=id; run;
```

```
Fit Statistics

-2 Res Log Likelihood         1546.7
AIC (smaller is better)       1550.7
AICC (smaller is better)      1550.7
BIC (smaller is better)       1554.9
```

c. Consider the same model as above, but include a spatial power covariance structure for repeated measures. Does the addition of the new $\mathbf{R}_i$ structure improve the model fit?

Solution: Yes, the AIC drops a couple of points after adding the SP(EXP) structure.

```
proc mixed data=long.Be_study;
class sex id;
model aado2r=ntep height sex ageep
   / solution;
random intercept / subject=id;
repeated / subject=id
type=sp(exp)(ntep); run;
```

```
Fit Statistics

-2 Res Log Likelihood          1542.8
AIC (smaller is better)        1548.8
AICC (smaller is better)       1548.9
BIC (smaller is better)        1555.1
```

d. Using relevant output and formulas (for the model in part c), estimate the correlation between $Y_{ij}|b_{0i}$ and $Y_{ij'}|b_{0i}$ for responses that are (i) 4 months apart, (ii) 1 year apart. Interpret the results for the study.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	id	12.2551
SP(POW)	id	0.04757
Residual		28.3927

Solution: here I am asking for elements of the R matrix (when conditioning on the random effects).

(i) $0.04757^{(4/12)} = 0.3623$
(ii) $0.04757^{1} = 0.04757$
Interpretation for (i): For a given subject and their deviation from the population mean, two measures 3 months apart have a correlation of 0.36.

e. Repeat part d, but for the correlation between Y_{ij} and $Y_{ij'}$ (unconditional responses).

Solution: here I am asking for elements of the V matrix (not conditioning on the random effects)

(i) $(28.3927*0.04757^{(4/12)} +12.2551) / (28.3927+12.2551) = 0.5546$
(ii) $(28.3927*0.04757+12.2551) / (28.3927+12.2551) = 0.3347$
Interpretation for (i): Two measures within a subject that are 3 months apart have a correlation of 0.55 (not conditioning on their random deviation from the mean).

f. Include a streamlined version of the output and write a few sentences of summary of the results. Make sure to include an estimate the progression of the illness (slope of *ntep* term), with a 95% confidence interval. For $AADO_2R$, remember that the lower the value, the better the health. What can be concluded about the progression of the illness over time?

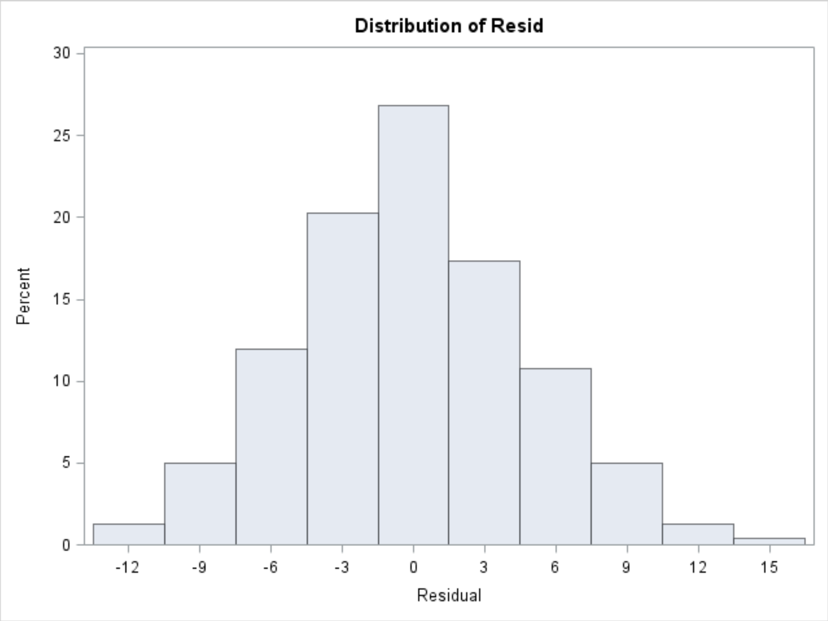
Solution: By adding CL as an option in the MODEL statement, we can get confidence limits. Below are parameter estimates in the model, for both fixed effects and covariance parameters. There is an average increase of 0.2117 units in AADO2R (units are mmHg) for each additional year with the disease, after removing effects of aging (95% CI: 0.055 to 0.368). SAS output follows on the next page.

```
Covariance Parameter Estimates

Cov Parm      Subject    Estimate
Intercept     id          12.2621
SP(EXP)       id           0.3283
Residual                  28.3902

Solution for Fixed Effects
                          Standard
Effect     sex  Estimate     Error    DF  t Value  Pr > |t|   Alpha     Lower     Upper
Intercept        -3.3030   14.5816    58    -0.23    0.8216    0.05  -32.4912   25.8852
ntep              0.2117   0.07919   179     2.67    0.0082    0.05   0.05544    0.3680
height            0.1326    0.2214   179     0.60    0.5501    0.05   -0.3043    0.5694
sex        1     -2.1333    1.8690   179    -1.14    0.2552    0.05   -5.8213    1.5547
sex        2           0         .     .        .         .       .         .         .
ageep           0.007743   0.07610   179     0.10    0.9191    0.05   -0.1424    0.1579
```

g. Note that $AADO_2R$ actually is a continuous variable with a clump at 0. The best model for these data might actually be a mixture model (to be discussed later, if time). However, examine a histogram of the residuals to determine whether the current model may be acceptable. Comment on what you see. [Aside: in this case, it is likely that the 0's should have been positive since a 0 value does not appear to be theoretically possible for this variable; from what I can see they can be attributed to measurement error, which may have affected other values as well.]



Solution: The residuals are pretty normally distributed! So, despite having a mixed distribution (continuous plus 0's), it seems that we've satisfied distributional assumptions for the model. Still, we can use mixture model techniques for these data, to be discussed soon. Also, a more thorough investigation would involve reviewing other model diagnostics, not just a histogram of the residuals…