Homework 4 Statistics 8320 (Wikle) March 8, 2016

DUE: March 17, 2016

1. Consider the model $Y_{ij} = \beta_0 + b_{0i} + b_{1i}W_{ij} + \epsilon_{ij}$, for $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, where
$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix} \right), \quad \epsilon_{ij} \sim \ iid\ N(0, \sigma^2).$$
   (i) Assuming that $n_i = 2$ and $W_{i1} = 1$, $W_{i2} = 2$ and $\sigma^2 = 2$, what is the marginal variance/covariance matrix of $\mathbf{Y}_i$?
   (ii) Assuming that $n_i = 2$ and $W_{i1} = 1$, $W_{i2} = 2$ and $\sigma^2 = 2$, what is the conditional variance/covariance matrix of $\mathbf{Y}_i$?
   (iii) This model was fit on data giving a value of -2 times the log-likelihood of 420 with 14 degrees of freedom. In addition, we considered the model in which $\text{cov}(b_{0i}, b_{1i}) = 0$, which gave a value of -2 times the log-likelihood of 426. Test (at $\alpha = .05$) whether we should favor the model for which the random effects parameters are dependent.

2. Suppose
$$Y_t = \beta x_t + t\, b_t + \epsilon_t, \ \ \epsilon_t \sim \ iid\ N(0, \sigma^2), \ \ t = 1, 2, \ldots$$
where $b_t \sim N(0, \tau^2)$ and
$$corr(b_t, b_{t-k}) = \begin{cases} 1, & k = 0 \\ \frac{\phi}{1+\phi^2}, & k = \pm 1 \\ 0, & \text{otherwise.} \end{cases}$$
Assume we observe y_t and x_t for $t = 1, 2$.
   (a) Write the model in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\epsilon}$ where $cov(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$ and $cov(\mathbf{b}) = \mathbf{D}$. Report $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{b}$, $\mathbf{Z}$, $\boldsymbol{\epsilon}$, $\mathbf{D}$ and $\boldsymbol{\Sigma}$ for this specific model.
   (b) What is the marginal variance of Y_2 and the marginal covariance between Y_1 and Y_2? (Write these in terms of the parameters τ^2, ϕ, and σ^2.
   (c) Assume that the estimates for the τ^2, σ^2, and ϕ parameters in the above model were estimated in SAS via REML. The model presented above was fit and the value of -2 times the log-likelihood was found to be 200. Then, another covariate was added, say X_2 and the model was fit via REML and the value of -2 times the log-likelihood was found to be 197. What would you learn from a likelihood ratio test in this case?

3. The growth pattern of an experimental soybean variety was studied. Eight plots were seeded and the average leaf weight per plot was assessed at weekly intervals following germination. If t measures time since seeding in days, the average growth is assumed to follow a linear model:
$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + e_{ij}, \tag{1}$$

where Y_{ij} is the average leaf weight per plant on plot i measured at time t_{ij}. The double subscript for the time variable t allows measurement occasions to differ among plots. It was decided to add a random effect so that the model is written:

$$Y_{ij} = \beta_0 + (\beta_1 + b_{1i})t_{ij} + \beta_2 t_{ij}^2 + e_{ij}, \tag{2}$$

where $b_{1i} \sim \ iid\ N(0, \sigma_b^2)$ and $e_{ij} \sim \ iid\ N(0, \sigma^2)$.

(a) If you graphed leaf weight per plant versus time (days after seeding) for each of the eight plots, what would you expect to see that would suggest the addition of the random component given in model 2?

(b) Write model 2 in matrix notation $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$ and define: $\mathbf{X}_i$, $\mathbf{Z}_i$, $\boldsymbol{\beta}$, $\mathbf{b}_i$, and $var(\mathbf{e}_i)$. Assume there are n_i times associated with the i-th plot.

(c) Derive the marginal variance-covariance matrix $var(\mathbf{Y}_i)$ for model 2.

(d) What is the correlation between any two leaf weight measurements on the same plot in model 2? Assume $\sigma^2 = \sigma_b^2 = 1$ and let the repeated measurements in time be coded $t_{i1} = 1, t_{i2} = 2, t_{i3} = 3$, etc. Show that the correlations do not decrease with temporal separation. Is this result realistic?

(e) In this problem, discuss why it might be advantageous to model the marginal covariance derived hierarchically as above versus modeling the marginal covariance, say $\boldsymbol{\Sigma}_i$, directly, where $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$. In this case, there are no *a priori* assumptions about the structure of $\boldsymbol{\Sigma}_i$.

4. An experiment was conducted in order to investigate four different treatments of pasture and two mineral supplements on milk yield. Twenty-four cows were used in the experiment and the experiment was designed as a split-plot, with pasture treatments (factor A) assigned to the main plots and mineral supplements (factor B) assigned to split-plots. The experiment was replicated in three blocks. The data can be read into SAS from the commands following the question below.

(a) Find the ANOVA table corresponding to this analysis. Use this table to infer which effects are important.

(b) What does the following SAS statement mean if it is in the MODEL line "DDFM = KENWARDROGER"? (Hint: look at the SAS manual). Does it change the results to have this?

(c) Consider the differences in least squares means, using a Tukey adjustment for multiple tests. Which pasture differences are significant?

```
DATA split;
   INPUT plot past min milk @@;
   DATALINES;
    1 4 2 30    1 4 1 29    1 1 2 27    1 1 1 25
    1 2 1 26    1 2 2 28    1 3 2 26    1 3 1 24
    2 2 1 32    2 2 2 37    2 1 2 30    2 1 1 31
    2 4 1 34    2 4 2 37    2 3 1 33    2 3 2 32
    3 1 2 34    3 1 1 31    3 2 1 30    3 2 2 31
    3 4 2 36    3 4 1 38    3 3 1 33    3 3 2 32
 ;
```

5. Alzheimer's disease is a condition that involves progressive deterioration in all aspects of intellect, self-care, and personality. Recent work suggests that the disease invovles pathological changes in the central cholinergic system, which might be possible to remedy by long-term dietary enrichment with lecithin. It is hoped that such treatment might slow or perhaps even halt the memory impairment associated with the disease. A study was conducted in which patients suffering from Alzheimer's disease were randomly allocated to receive either lecithin or a placebo for a 6-month period. A cognitive test score giving the number of words recalled from a previously given standard list was recorded monthly for 5 months. The main question of interest is whether the lecithin treatment had any effect.

   Consider the data in `alzheim.dat` on the class website. The first column of data indicate the *Group*, with 1 indicative of the placebo group and 2 the lecithin group. The following 5 columns contain the monthly scores in sequence. Note, you can read the data in the following fashion, with one case per measurment. The grouping variable and five monthly scores for each subject are read in together, and then split into separate observations using the "array", iterative do loop, and "output" statement. Note, the "visit" variable is assigned the values 1,2,3,4,5, depending on the month of the test. With 47 subjects and 5 visits each, the resulting data set contains 235 observations.

```
data alzheim;
  infile 'alzheim.dat';
  input group score1-score5;
  array sc {5} score1-score5;
  idno=_n_;
  do visit=1 to 5;
    score=sc{visit};
    output;
  end;
run;
```

   Note, you may also want to sort the data by group to facilitate plotting.

   (a) Plot the profiles for each subject (score on y-axis and "visit" on x-axis). I recommend plotting each group separately (or the plots get to "messy"). Describe the similarities and differences in profiles for each group. (attach your plot)

   (b) Let y_{ijk} represent the cognitive score for subject k on visit j in group i. Write out a mixed-effects model that regresses these scores on `visit` (this variable takes the values 1,2,3,4,5) and `group` and allow a random intercept. You may assume normal, independent errors.

   (c) Based on your model in part (b), what is the implied covariance for the repeated measures on each subject?

   (d) Fit the model from part (b). Use the maximum likelihood option for fitting (`method=ml`).

      i. What are your estimates for the variance components?

ii. Are the fixed effects significant. Interpret these results relative to the problem at hand (is there evidence that the treatment has helped?).

iii. Plot the predicted profiles for each subject by group. What do these plots suggest relative to the fit of the model?

(e) Write a mixed-model that has both a random intercept and a random coefficient for the `visit` variable.

(f) Fit the model in (e) using maximimum likelihood and an unstructured covariance for the random effects.

i. Write out the variance component estimates (including the covariance estimates). Are they significant? What is the estimated correlation between intercepts and slopes?

ii. Are the fixed effects significant? Interpret these results relative to the question to be addressed by this experiment.

iii. Is this model a better fit than the random intercept model? Explain.

iv. Plot the predicted profiles for each subject by group. How do these profiles compare to those from the random intercept model and the observed data?

(g) Assume that the conditional independence assumption for the repeated measures is no longer appropriate. Refit the random slope/intercept model but allow the errors to have an AR(1) structure (use the `REPEATED` option). Describe the results of your analysis relative to the iid error case from above. Try allowing an unstructured covariance model. Use AIC to decide which model is most appropriate.