Clustered data models - Exercises 3

1. When studying characteristics that affect student performance on a battery of exams, we may use a model that takes into account variation between schools, students and tests. The model for the score of student i in school s on test t is

$$y_{ist} = \mathbf{x}'_{ist}\boldsymbol{\beta} + u_s + v_t + w_{is} + \epsilon_{ist},$$

where $\{u_s\} \sim N(0, \sigma_s^2)$ is the school effect, $\{v_t\} \sim N(0, \sigma_t^2)$ is the test effect, $\{w_{is}\} \sim N(0, \sigma_w^2)$ is the student effect and $\{\epsilon_{ist}\} \sim N(0, \sigma_\epsilon^2)$ is the random error. All the random effects are assumed to be independent. This is a slight extension to the model introduced in Section 9.2.2 of Agresti.

Determine a) the intraclass correlation between scores on different exams for a student, b) the intraclass correlation between scores on a particular exam for a pair of students in the same school c) the intraclass correlation between scores on different exams for a pair of students in the same school.

2. (9.6 in Agresti) A crossover study comparing two drugs observes a continuous response (y_{i1}, y_{i2}) for each subject for each drug. Let $\mu_1 = \mathrm{E}(y_{i1})$ and $\mu_2 = \mathrm{E}(y_{i2})$ and consider $H_0 : \mu_1 = \mu_2$.

a) Construct the normal linear mixed model that generates a paired-difference t test (with test statistic $t = \sqrt{n}\bar{d}/s$, using mean and standard deviation of the differences $\{d_i = y_{i2} - y_{i1}\}$ and the corresponding confidence interval for $\mu_2 - \mu_1$.

b) Show the effect of the relative sizes of the variances of the random error and random effect on $\mathrm{Cor}(y_{i1}, y_{i2})$. Based on this, to compare two means, explain why it can be more efficient to use a design with dependent samples than with independent samples.

3. (9.8 in Agresti) For the extension of the random-intercept linear mixed model (9.8 in Agresti; page 51 in lecture slides) that assumes $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \sigma_\epsilon^2 \rho^{|j-k|}$, show that

$$\mathrm{Cor}(y_{ij}, y_{ik}) = (\sigma_u^2 + \rho^{|j-k|}\sigma_\epsilon^2)/(\sigma_u^2 + \sigma_\epsilon^2).$$

4. (9.32 in Agresti) For the smoking prevention and cessation study (Section 9.2.3 in Agresti; page 54 in lecture slides), fit multilevel models to analyze whether it helps to add any interaction terms. Interpret fixed and random effects for the model that has a SC $\times$ TV interaction.

5. (9.33 in Agresti) Using the R output shown for the simple analyses of the FEV data in Section 9.2.5, show that the estimated values of $\mathrm{Cor}(y_{i1}, y_{i2})$ and $\mathrm{Cor}(y_{i1}, y_{i8})$ are 0.74 for the random intercept model and 0.86 and 0.62 for the model that also permits autoregressive within-patient errors.

6. (9.34 in Agresti) Refer to Exercise 1.21 (in Agresti) and the longitudinal analysis in Section 9.2.5. Analyze the data in file FEV2.dat at www.stat.ufl.edu/~aa/glm/data, investigating the correlation structure for the eight FEV responses and modeling how FEV depends on the hour and the drug, adjusting for the baseline observation. Take into account whether to treat hour as qualitative or quantitative, whether you need interaction terms, whether to have random slopes or only random intercepts, and whether to treat within-patient errors as correlated. Interpret results for your final chosen model. (You may want to read Littell et al. (2000). The book SAS for Mixed Models, 2nd ed., by Littell et al. (2006, SAS Institute), uses SAS to fit various models to these data.)