STAT 8630 — Mixed-Effect Models and Longitudinal Data Analysis
Final Exam – Due Thursday, May 2
SHOW ALL WORK

Name:_____

Do not discuss this exam with anyone other than your instructor. You may use any notes, books, journal articles etc., but you are not to collaborate!

1. The dataset tvsfpors.dat on the course website contains data from the Television School and Family Smoking Prevention and Cessation Project (TSFP). In this study, schools were randomized to one of four conditions given by the combinations of of the levels of two experimental factors: (1) TV, a television intervention (1=present, 0=absent), (2) CC, a social-resistance classroom curriculum (1=present, 0=absent). The outcome measure is the tobacco and health knowledge scale (THKS) score, an instrument designed to measure tobacco and health knowledge. This variable has been collapsed into four ordinal categories.

   In this study design, students are nested in classes, which are nested within schools. It was desired to generalize the results from this study to a broader population of children from all classes and all schools in the target population. The dataset tvsfpors.dat has the following variables:

   - **school:** school identifier
   - **class:** class identifier
   - **thk:** ordinal THKS score measured postintervention (the response variable)
   - **prethk:** ordinal THKS score measured preintervention
   - **cc:** social-resistance classroom curriculum (0-1 variable)
   - **tv:** television intervention (0-1 variable)

   a. In this problem we will fit models of the form $\mathbf{y}_{ij}|b_i \sim Mult(1, \boldsymbol{\pi}_{ij}^c)$ where $\mathbf{y}_{ij}$ is a $q \times 1$ vector of indicators for which level is observed for a $q+1$-category response variable z_{ij}. We assume that conditional on the b_is, $\{\mathbf{y}_{ij}\}$ are independent over i and j. We assume a cumulative logit model of the form

   $$\text{logit}(\gamma_{ijr}) = \alpha_r + \mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i, \quad r = 1, \ldots, q, \qquad (*)$$

   where $b_1, ..., b_n \overset{iid}{\sim} N(0, \psi)$ and $\gamma_{ijr} = \sum_{k=1}^{r} \pi_{ijk}$. Show how this model can be written as a threshold model for an underlying latent random variable z_{ij}^* that follows a LMM of the form

   $$z_{ij}^* = -\mathbf{x}_{ij}^T\boldsymbol{\beta} - b_i + \epsilon_{ij}$$

   where, conditional on the b_is, $\{\epsilon_{ij}\} \overset{iid}{\sim}$ logistic, with cumulative distribution function $F(\tau) = \frac{e^\tau}{1+e^\tau}$. Show that the intraclass correlation coefficient for the latent response $\rho \equiv \text{corr}(z_{ij}^*, z_{ik}^*)$ is given by $\rho = \psi/(\psi + \pi^2/3)$ and

explain why within-cluster correlation is more often quantified in terms of $\text{corr}(z^*_{ij}, z^*_{ik})$ rather than $\text{corr}(z_{ij}, z_{ik})$ or some sort of multidimensional correlation between $\mathbf{y}_{ij}$ and $\mathbf{y}_{ik}$.

b. Investigate how tobacco and health knowledge is influenced by the interventions by fitting a proportional odds model of the form (*). Include main effects for **cc**, **tv**, and the interaction between these two factors. In addition, control for **prethk** appropriately. Include random effects for schools, but ignore any possible effects of class. As part of your analysis, do the following:

i. Obtain estimated odds ratios for all fixed effects in your model and interpret them. Also compute and interpret approximate 95% confidence intervals for the corresponding odds ratios.
ii. Estimate the intraclass correlation ρ for within-cluster correlation in the latent responses.

c. Now refit your model from part b, but with random effects for class instead of school. That is, in part (b) we treat students as nested within schools (ignoring class), here we treat students as nested within class (ignoring schools). As part of your analysis, do the following:

i. Obtain estimated odds ratios for all fixed effects in your model and interpret them. Also compute and interpret approximate 95% confidence intervals for the corresponding odds ratios.
ii. Estimate the intraclass correlation ρ for within-cluster correlation in the latent responses.
iii. Comparing your results from parts (b) and (c), does there appear to be more within-cluster correlation within schools or within classes? Is this result intuitively reasonable? Explain.

d. Now fit a three level proportional odds model with random effects for schools and for classes within schools and the same fixed effects as in the models from parts (b) and (c). Write down the model that you are fitting in its general mathematical form including all assumptions and giving all details of your notation. Then, based on the fitted model, do the following:

i. Using appropriate statistical methodology, determine which of the three models is most appropriate for these data. Justify your choice.
ii. Compare the variance components from the three level model fit in part (d) to the variance component due to class from part (b). What is the approximate relationship between them, and why is this relationship intuitively reasonable?
iii. Describe the fitting methodology you used and any computational problems that came up in fitting the models from parts (b)–(d) of this exercise. Explain how you dealt with those problems.

2. In Don Hedeker's notes that we covered in class, he gave formulas for the pattern mixture averaged parameter estimator $\hat{\bar{\boldsymbol{\beta}}}$ (p.20) and an approximate variance estimator for this quantity, $\hat{V}(\hat{\bar{\boldsymbol{\beta}}})$ (p.29). These estimators were given and illustrated in the case where the pattern mixture model involves just two categories for the dropout variable: completers and non-completers. Now consider the case where the dropout variable is a $q+1$ -level categorical where $q > 1$. In that case, the formula for $\hat{\bar{\boldsymbol{\beta}}}$ generalizes in an obvious way:

$$\hat{\bar{\boldsymbol{\beta}}} = \hat{\pi}_c \hat{\boldsymbol{\beta}}_c + \hat{\pi}_{d_1} \hat{\boldsymbol{\beta}}_{d_1} + \cdots + \hat{\pi}_{d_q} \hat{\boldsymbol{\beta}}_{d_q}$$

where $\hat{\pi}_c, \hat{\pi}_{d_1}, ..., \hat{\pi}_{d_q}$ are the sample proportions of subjects who are completers, dropouts of type 1, dropouts of type 2,..., dropouts of type q, respectively (here I am assuming that dropouts of type q+1 are completers). In addition, $\hat{\boldsymbol{\beta}}_c, \hat{\boldsymbol{\beta}}_{d_1}, ..., \hat{\boldsymbol{\beta}}_{d_q}$ are the parameter estimates of interest for the different strata defined by the dropout variable.

   a. In the case where the dropout variable has > 2 categories, derive a formula for $\text{var}(\hat{\bar{\boldsymbol{\beta}}})$ in terms of $\text{var}(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{\pi} = \text{E}(\hat{\boldsymbol{\pi}})$ where $\hat{\boldsymbol{\pi}} = (\hat{\pi}_c, \hat{\pi}_{d_1}, ..., \hat{\pi}_{d_q})^T$. In addition, propose an estimator of $\text{var}(\hat{\bar{\boldsymbol{\beta}}})$.

   b. Use the results from part (a) to obtain the pattern-mixture averaged results corresponding to the model on p.16 of Hedeker's notes (the parameter estimates from this model are in the last column on p.17 of those notes). Be sure to obtain pattern-mixture averaged estimates of the parameters of interest and corresponding standard errors. Compare your results with those from the ordinary and pattern-mixture averaged results given on p.33 of Hedeker's notes. The latter are based on a more restrictive pattern-mixture model with just two categories of missingness: completers and dropouts. Summarize and interpret your results in the context of the dataset and the corresponding study of schizophrenia and its treatment.

3. Expression (13.3) in our text gives the sandwich estimator in a marginal model for longitudinal data y_{ij}, $i = 1, ..., N$, $j = 1, ..., n_i$ when fit with the GEE methodology. Also given is the model-based variance-covariance estimator, which takes the form $\mathbf{B}^{-1}$ in the authors' notation, and the asymptotic variance-covariance estimator that does not assume that the model was fit under a correct specification of $\text{var}(\mathbf{y}_i)$ (this one takes the form $\mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1}$). Consider the special case when $n_i = 1$ for all i and let $y_i \equiv y_{i1}$ for all i. Suppose we have count data $y_1, ..., y_N$ which we assume are independent with $\text{E}(y_i) = \mu_i$ where $\mu_i = \beta$ (an identity link model with constant mean).

   a. Suppose we fit this model under a working assumption that $\text{var}(y_i) = \mu_i$, as is the case for Poisson data, but in truth, the variance of y_i is μ_i^2. Show that the model-based variance of the GEE estimator $\hat{\beta}_{\text{GEE}}$ is β/N and its estimator is $\bar{y}/N$. In addition, show that the sandwich estimator of $\text{var}(\hat{\beta}_{\text{GEE}})$ is $\sum_i^N \frac{(y_i - \bar{y})^2}{N^2}$. Which of these two estimators would you

expect to be better (i) if the Poisson model holds, and (ii) if there is severe overdispersion relative to a Poisson model? Explain.

b. Now suppose we assume that $\text{var}(y_i) = \sigma^2$ when in truth $\text{var}(y_i) = \mu_i$. Find the model-based asymptotic variance, the actual asymptotic variance, and the sandwich estimator of the actual variance for $\hat{\beta}_{\text{GEE}}$.