STAT/BIOST 571: Homework 8

Philip Pham

March 8, 2019

Problem 1: GEE and GLMM; interpretation of marginal parameters in logistic regression models; missing data (20 points)

Download the `fluoride.csv` *dataset from the course website. This dataset contains 3846 observations of fluoride intake for 1279 children, with follow-ups at ages 1.5, 3, 6, and 9 months, but with some observations missing for individual children. The variable id indexes unique children, age denotes age in months, income is an indicator for maternal income over 30 thousand dollars per year,* $fluoride$ *is total fluoride intake (mg per kg of body weight), and* fl *is an indicator for* $fluoride > 0.05$*. Our primary interest is the relationship between the binary outcome* fl *and the child's age, potentially including effect modification by maternal income. We will fit logistic regression models for the* fl *outcome with the standard mean variance relationship and either a multiplicative interaction*

$$\mu = expit(\beta_0 + \beta_1 \times age + \beta_2 \times income + \beta_3 \times age \times income) \tag{1}$$

or just an intercept and a main effect

$$\mu = expit(\beta_0 + \beta_1 \times age). \tag{2}$$

In all analyses, we account for correlation within children and assume the data from different children are independent.

(a) *Fit model (1) using GEE with independence and exchangeable working correlation models and using a standard GLMM with random intercepts. Report point estimates and standard error estimates for all four regression coefficients and all three model fits in a single table (use robust standard errors for GEE and model-based versions for GLMM).*

(b) *Discuss any differences between the estimated values of* β_1 *from your three fitted models.*

(c) *For each of your three fitted models, write a short paragraph summarizing your main findings. Specifically, give scientifically interpretable statements (including confidence intervals) about the relationships between fluoride intake and age in children with maternal income greater than 30 thousand dollars per year and in children with maternal income less than 30 thousand dollars per year.*

(d) *Now repeat part (a), but use model (2) instead of (1) (there are now only two regression coefficients to report per model).*

(e) *Download the dataset* `fluoride.miss.csv` *from the course website and repeat the calculations from part (d). Note* `fluoride.miss.csv` *is a subset of* `fluoride.csv`*, with more missing data.*

(f) *Discuss the differences between your results in parts (d) and (e). Speculate about the missingness mechanism that gave rise to the* `fluoride.miss.csv` *dataset and explain how this might account for what you observe. You might find it helpful to conduct exploratory analyses of the two datasets and to consider your findings from part (a) of this problem.*