HOMEWORK 6
BIOSTATISTICS 755

This homework can be handed in any time up to 5pm on April 27th, 2023

1. **(50 points total)** This question is going to use the Study of Assets and Health Dynamics among the Oldest Old (AHEAD) data that we used in the previous homework. Please see that assignment for a description of the study objectives and variables. Here, the response variable will be again be `iadlany,` an indicator for whether or not the subject reports difficulty with any (at least one) IADL. The main variable of interest is total word recall (TWC). The AHEAD data set has considerable missing data. As this population is older, let's assume that the main reason for dropping out of the study is death.

   The GEE model fit in the solutions to the homework was:

$$\log\left\{\frac{P(IADLANY_{ij}=1)}{P(IADLANY_{ij}=0)}\right\} = \beta_0 + \beta_1 AGE_i + \beta_2 TWR_{ij} + \beta_3 TWR_{ij}^2 + \beta_4 SEX_i \quad (1)$$

   please use this form for the fixed effects.

   In the dataset `ahead_mono.csv` I've added two important variables for this analysis. First, a lag variable `iadlany_lag` which will have the previous `iadlany` value. This variable can only be defined for years 2, 5, and 7. Also, the data contain `iadlany_mono` which is the `iadlany` changed so that the missingness is monotone. The `iadlany_mono` version of the `iadlany` variable is to be used only for part (d) only.

   Note that this data have 26604 records. The large size of this dataset is part of the difficulty of this problem. You are very likely to come across big annoying datasets sometime in your career. If you have been using SASOnline these data might cause issues (i.e., won't run) due to the large memory demands. The desktop version of SAS is available in the 4th floor computer lab where memory issues won't be an issue. A smart strategy in situations like this is to first limit the data to a smaller subset (maybe the first 1000 records) and get all your code working/debugged using the smaller version. Once your code is finalized, you can rerun on the larger dataset (possibly in the computer lab if your having issues with your PC). If your code is correct, you only have to run on the large dataset once.

   (a) **(10 points) Qualitatively** assess whether the missing data mechanism is MCAR, MAR or MNAR. Be as specific as possible using the vocabulary we discussed in class.

   (b) **(10 points)** Regardless of your assessment in (a), **quantitatively** if the data are MCAR. To do this use the missing data indicator (`MissInd`). Fit the following logistic GLMM with a random intercept using the `MissInd` variable as the outcome and `iadlany_lag` as the only predictor.

$$\log\left\{\frac{P(MissInd_{ij}=1)}{P(MissInd_{ij}=0)}\right\} = \beta_0 + b_{i0} + \beta_1 IADLANY\_lag_{ij}$$

Is the β_1 coefficient significant? How is the lag iadlany variable related to missing an observation? What does that indicate about the MCAR assumption?

(c) **(15 points)** Perform an analysis of the model in equation (1) using multiple imputation for the missing data in the model (only for those variables used in the model) with a generalized linear mixed model with a random intercept.

(d) **(15 points)** Fit a model with Age, sex and Age*Sex with **weighted and unweighted logistic GEE**. Report the parameter estimates for the weights (i.e., the `mismodel` parameter estimates) and the regression model. To do this you'll need to use the monotone version of the outcome.

2. **(50 points total)** The data `school_math.xlsx'` on the course website is a dataset with the following variables:

- `Classid`: class identifier
- `Childid`: Student identifier
- `Schoolid`: school identifier
- `SES`: student's socioeconomic status
- `Housepov`: average household poverty by school
- `Mathprep`: class's amount of time spent on math prep
- `sex`: male=1
- `Mathknow`: class's mean math score
- `Yearstea`: years teacher has taught
- `minority`: minority=1
- `mathkind`: math score in kindergarten
- `mathgain`: points gained in math score since kindergarten

which are measured on multiple levels. Use this data to answer the following questions.

(a) **(5 points)** How many levels does the data have? Which variables are measured on which level?

(b) **(10 points)** Using mathkind as the outcome, calculate the unadjusted ICC at ***all*** levels. Give an interpretation of both ICC values.

(c) **(5 points)** Using the previous analysis, calculate the plausible values range of school level mean kindergarten math scores.

For the remaining questions use mathgain as the outcome.

(d) **(5 points)** Calculate the unadjusted ICC at ***all*** levels. Give an interpretation of both ICC values.

(e) **(10 points)** Which level 2 variable results in the biggest *decrease* is the level 2 ICC when it is adjusted for? Report the % decrease of this variable.

(f) **(5 points)** What is the change in the level 3 ICC when the level 3 variable is adjusted for? Report the % decrease of this variable.

(g) **(10 points)** Does the impact of student SES vary across **schools**? Run a model that adjusts for student SES, and test if there's heterogeneity in the impact of SES at the school level. That is, fit the following model (here i is student, j is the class, and k is school):

$$\begin{aligned} mathgain_{ijk} &= \beta_{0jk} + \beta_{1j} SES_{ijk} + e_{ijk} \\ \beta_{0jk} &= \gamma_{00k} + b_{0jk} \\ \gamma_{00k} &= \gamma_{000} + b_{0k} \\ \beta_{1j} &= \gamma_{10} + b_{1k} \end{aligned}$$

where $\binom{b_{0k}}{b_{1k}} \sim N(\mathbf{0}, \boldsymbol{G})$ and $b_{0jk} \sim N(0, \sigma_b^2)$. Report the results here.