STA/BST 224 Longitudinal Data Analysis

Final Problem
You have 7 days (24×7 hours) to work on this final problem

Instruction:

- This final problem set is to be done **alone**, without discussing it with or consulting any other students in the class, or anyone else except for the instructor or the course assistant. You are allowed to use the notes, the text and other texts as you wish. However, you are not allowed to consult any specific solutions to this problem that you might find on the web.
- You are not required to write a formal report for it, just need to submit your solution like previous homework. (So it is more like a homework than a project.)
- In case you are more interested in analyzing your own data, please contact instructor first to make sure it is ok to use it as a substitute for this problem. Also, you need to add a paragraph about the relevant background, the motivation of your study, and describe the outcome and covariates variables. Since the lecture about missing data will be in last class, you are ok to simply ignore missing values in data.
- Please submit it through Canvas. You can scan or take picture of hand-written solution as long as it is clear (merge all files into one).
- You can use either R/SAS/Stata or other software. Please include program and important results.

1. A study (ICHS) was conducted in West Java, Indonesia, to determine the effects of Vitamin A deficiency in preschool children. The data file (ichs.csv) is available on the course web page. The goal of this problem is to analyze these data using marginal model (GEE) to address the interests of the investigators. The investigators were particularly interested in **whether children with Vitamin A deficiency were at increased risk of developing respiratory infection**, which is one of the leading causes of death in this part of the world. Of course, any association between Vitamin A deficiency and respiratory infection could be confounded by age and/or gender of the child, and the effect of Vitamin A deficiency on risk of respiratory infection may also vary by age and/or gender of the child.

   Two-hundred fifty children were recruited in the study, and their (baseline) age in years (`bage`), gender (`gender`: 0 = male, 1 = female), and whether they suffered vitamin A deficiency (`vita`: 0 = no, 1 = yes) was recorded at an initial clinic visit (time 0). Also recorded was the response, whether the child was suffering from a respiratory infection (`infect`: 0 = no, 1 = yes). The children were then re-examined at 3 month intervals for 15 months (at 3, 6, 9, 12, and 15 months) after the first visit, and the presence or absence of respiratory infection was recorded at each of these visits. Luckily (for you), all children were seen at all visits, so there are no missing data in this study.

   Students can choose appropriate tools (as long as your methods are well justified in your solution), and are not required to do the exactly same analysis as the solution provided by instructor.

   (a) Do an exploratory analysis of this data set: provide some tables and/or graphics to describe the covariates and response; explore the data with respect to the primary scientific aim of the study; provide a brief description summarizing the results. For example, you can report distribution of baseline covariates across subjects, distribution of the response variable for different time and groups, explore how response changes by checking how many subjects with an infection at a given wave will still have an infection at the following wave, etc.

   (b) Continue to do exploratory analysis for correlation structure of the response variable. Choose an appropriate flexible generalized linear model (ignore correlation for now for exploration purpose). Write down your model: linear predictor (and what covariates will be included in it), link function,

and the variance function for this model. The model should be flexible (eg, you can include all two-way interactions). Then explore the correlation structure of the response variable using whatever tools you feel are appropriate. From this exploration, decide on a working correlation structure to be used in modelling these data.

(c) Using GEE with robust variance estimator, and the working correlation structure you chose, fit a series of marginal models with the goal of finding a single model to address the investigators' questions of interest. You might start with your flexible model in (b) or other initial model, and consider interaction terms, removing unimportant terms, etc. Clearly describe your sequence of model fits, hypothesis tests along the way, and thinking about including or dropping terms. Present your final model fit, including confidence intervals for each parameter for fixed effects. Also provide an estimate of correlation model parameters fitted in your model

(d) Confirm the mean response is well-captured by the fitted mean model, by comparing your model fit to the empirical proportion of subjects with respiratory infections. (For example, you may check linearity of time and baseline age, by plotting the empirical and predicted proportions by vitamin A deficiency and time (or baseline age), where predicted proportion can be the average of your fitted values.) (Hint: to obtain predicted values, you may try `output out=predict pred=predict;` in SAS Proc genmod; or use `fitted.values` after obtaining fitted `geeglm` in R)

(e) Give a "marginal model" interpretation for some of the typical parameter estimates in your final model. (Note: You do not have to write out interpretations for **all** parameters in your model, but provide enough that the reader is convinced you know how to correctly interpret the model.) Also explain the estimate of correlation model parameters fitted in your model. Summarize your findings in a brief paragraph. What do these data say about the effect of Vitamin A deficiency on risk for respiratory infection?

(f) Fit a conditional model (GLMM with a random intercept), using the covariates chosen in previous final marginal model. Interpret the estimated coefficients (for fixed effects) and variance parameter (for random intercept). Compare the coefficient estimates from the conditional model and marginal model. Are their differences (or ratios) in agreement with what the theory predicts?