PUBH 7430

Assignment 4

On this assignment, as in all assignments, **UNEDITED COMPUTER OUTPUT IS NOT ACCEPTABLE**. *Tables and plots prepared from statistical output should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. Obtaining the correct results from an analysis is only a portion of each exercise. It is also very important to write clearly about what your results mean. This includes presenting your results in a way that clearly answers the question and places little burden on the reader.*

You should submit your assignment on Canvas as a PDF file, titled "PubH7430_LastNameFirstName_A4.pdf" with your own name replacing LastNameFirstName. While no code should be included in the body of the assignment, you **must include your code as an appendix** *at the end of the assignment. Keep in mind that while working together on homework assignments is permitted, each student is expected to independently write up homework assignments, including any code, in their own words.*

Questions

1. **Effect of time on vitamin E levels using LMM [16 pts].** For this question, we will be using data from a longitudinal Phase II randomized trial evaluating the effect of beta-carotene supplementation on vitamin E levels. We will only use data from the group who received no supplementation (i.e., the control group). Each subject has outcomes measured at three time points: baseline, 3 months, and 9 months. You should treat all three of these outcomes (including the baseline measurement) as a set of correlated outcomes for each subject. Also, you should treat time (in months post-baseline) as a continuous variable.

   The data are available in long format on Canvas, saved as `vitaminE.csv`. The columns are as follows:

   - **id:** Patient ID
   - **time:** Visit number (values 1, 2, or 3 with 1 = baseline)
   - **timemonths:** Timing of measurement in months post-baseline where 0 = baseline
   - **vite:** Vitamin E level in μg/dl

   NOTE: For this questions, you will want to calculate p-values and confidence intervals for fixed effects. As mentioned in class, inference on fixed effects in mixed models is a subject of current research. If you are using `SAS`, you can just use the default settings for calculating degrees of freedom. If you are using the `lme4` package in `R`, p-values will not be computed by default. I recommend that you use the `lmerTest` package to compute them.

   (a) [2 pts] Make a spaghetti plot of individual vitamin E trajectories over time.

(b) [2 pts] **Model A.** Fit a simple linear model of vitamin E level versus time, which ignores the correlation between outcomes on the same subject. Report the estimate and standard error for the time effect.

(c) [3 pts] **Model B.** Fit a linear mixed effects model for vitamin E level with a fixed effect for time (in months post-baseline) and a random intercept. Report the estimate, standard error, and 95% confidence interval for the time effect. In a sentence suitable for a scientific publication, interpret the time effect.

(d) [2 pts] How does the SE of the time effect from the simple linear model (Model A) compare to the SE of the time effect from the mixed model with random intercepts (Model B)? Comment on how this fits with what we have learned previously about the impact of ignoring correlation.

(e) [2 pts] For model B, report the estimated SD of the random intercepts and the estimated SD of the residual errors. What does this tell you about the amount of variability within subjects compared to between subjects?

(f) [3 pts] **Model C.** Now fit a linear mixed effects model for vitamin E level with a fixed effect for time along with random intercepts and *random slopes for time*. Report the estimate, standard error, and 95% confidence interval for the time effect. In a sentence suitable for a scientific publication, interpret the time effect.

R coding notes: When you fit this model in `R`, you may see the arning "boundary singular fit: see ?isSingular." You may ignore this for the purposes of this question.

(g) [2 pts] Perform a statistical test to establish whether the model with random intercepts and random slopes has significantly better fit than the model with only random intercepts, and report your results. What reference distribution is used for this test?

Note: Due to differences in model fitting, the results from `R` and `SAS` will be different. Just report the results that you find given your software choice.

2. **Treatment differences over time using GEE and GLMM [24 pts].** For this question, you will be analyzing data from a clinical trial in which 59 individuals with epilepsy were randomly assigned to either an anti-seizure medication (progabide, n=31) or placebo (n=28). Prior to treatment, study participants reported the number of seizures they experienced in the previous two weeks. They also reported the number of seizures they experienced during four consecutive 2 week periods after initiation of study drug (placebo or progabide). When fitting your models, you should treat all of these outcomes (including the baseline measurement) as a set of correlated outcomes for each subject. We will assume that the number of seizures during each two-week post-baseline period follows a Poisson distribution, i.e., we will model the seizure rate (mean # of seizures per two-week period). The goal of your analysis is to compare the effect of progabide vs. placebo on the changes in the seizure rate over the duration of the study.

   The data are available in log format on Canvas, saved as progabide.csv. The columns are as follows:

   - **subject**: Patient ID
   - **visit**: Visit number (values 0, 1, 2, 3, 4 with 0=baseline).
   - **trt**: Treatment where 1=progabide and 0=placebo
   - **seizures**: Number of seizures experienced in the previous 2 week period

(a) First, consider a *marginal* model for the rate of seizures.

   i. [2 pts] Write down the mean model for a GEE model for the rate of seizures. Allow each treatment group to have different time-varying treatment effects and different baseline seizure rates.

   ii. [2 pts] Fit the GEE model that you specified in 2(a)i assuming an exchangeable working correlation, and report estimates and 95% confidence intervals for all of the coefficients (including the intercept) on the scale that is most interpretable (i.e., exponentiate if needed). No interpretation is needed for this part.

   iii. [4 pts] Provide an interpretation for each of the estimates/CIs that you reported in 2(a)ii including the intercept

(b) Next, consider a *conditional* model for the patient-specific seizure rate.

   i. [2 pts] Write down the mean model for a generalized linear mixed model (GLMM) with **random intercepts** that assumes linear trends for the seizure rate over time, and allows for differences in baseline seizure rate between treatment groups. That is, you want to allow the treatment effect to vary over time in both treatment groups and possibly be different for the two treatments, and allow for the possibility of differences between the treatment groups at baseline.

   ii. [2 pts] Fit the GLMM model that you specified in 2(b)i and report estimates and 95% confidence intervals for all of the fixed effects (including the intercept) on the scale that is most interpretable (i.e., exponentiate if needed). No interpretation is needed for this part.
   Coding notes: In R, please include `nAGQ=50` as an argument in your `glmer()` function call so that `R` and `SAS` produce similar results. In SAS, please include `method=quad(qpoints=50)` as an option in `PROC GLIMMIX` so that `R` and `SAS` produce similar results.

   iii. [6 pts] Provide an interpretation for each of the estimates/CIs that you reported in 2(b)ii, including the intercept.

(c) [3 pts] Compare the estimates for the treatment by time interaction that you obtained from the conditional (GLMM) and marginal (GEE) models. Would we expect these terms to be the same or different?

(d) [3 pts] Which model (GLMM or GEE) do you think provides a more relevant estimate of the differences in the treatment effect over time in this setting? Provide a short justification. Note: Either may be appropriate if well-justified.