# PUBH 7430- Assignment 3

*On this assignment, as in all assignments,* **inclusion of screenshots of R/SAS 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_A3.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. **Time-varying treatment effect [28 pts].** In this question, you will be using the same dataset (*gumR* or *gumSAS*) as in Assignment 2. Please reference Assignment 2 for the description of the study and dataset. We will consider only the Xylitol gum (active treatment) and gum placebo arm (ignoring the "no gum" arm). We are interested in assessing whether the effect of treatment (Xylitol vs. placebo gum) on MS level varies over time. The outcomes of interest are the **post-baseline oral MS measurements**: *mutpost* (1 week post-baseline), *mut1mos* (4 weeks post-baseline), and *mut3mos* (12 weeks post-baseline). You will not be using the *mutpre* variable. You should analyze these as repeated longitudinal measurements (i.e., you should not analyze these three post-baseline measurements in separate models).In this question, we are interested in assessing whether the effect of treatment (Xylitol vs. placebo gum) on MS level varies over time.

   (a) [1 pt] Construct (and write down) a linear predictor which will allow you to assess whether the effect of treatment changes over time. Your time covariate should be measured in weeks since baseline and modeled as a continuous variable.

   (b) Using the linear model developed in 1 (a), fit a Gaussian GLM (thereby ignoring the correlation in the data) with identity link.

      i. [2 pt] Report the coefficient estimates with 95% confidence intervals for the GLM model you fit.

      ii. [2 pts] Are the coefficient estimates (i.e., $\hat{\beta}$) obtained an appropriate reflection of the mean mutans

streptococci levels for each treatment group over time? That is, are the coefficient estimates obtained trustworthy? Explain.

iii. [2 pts] How would you expect the standard errors for the coefficients in your GLM model to compare to standard errors in a GEE models? Explain how this fits with what we have learned previously about the impact of ignoring correlated outcomes in analyses.

iv. [2 pts] When fitting a Gaussian GLM, you could use the sandwich variance estimates to compute confidence intervals and p-values. Would this be likely to yield correct inference (i.e., confidence intervals and p-values) for these data? Why? Explain.

(c) Using the linear model developed in 1 (a), fit Gaussian generalized estimating equation (GEE) models using the identity link and the following working correlation matrices: independence, exchangeable, AR-1, and unstructured.

i. [5 pts] Summarize your results in a table listing the estimates with 95% confidence intervals for all coefficients. Comment on any similarities or differences in results for the different working correlations.

ii. [2 pts] In class, we saw that the results for independence and exchangeable working correlations were exactly the same for the Gaussian GEE. Why is this not the case here?

iii. [2 pts] For the GEE model with exchangeable working correlation, interpret the coefficient for the time by treatment interaction in a sentence or two suitable for inclusion in a scientific publication.

(d) Use the model generated in part (c) with exchangeable correlation to answer the following questions. For each question state what the mean value is in terms of the $\beta$ coefficients from 1(a) and provide point estimates and 95% confidence intervals.

i. [2 pts] What is the expected MS level 1 week after baseline for a participant who received the treatment gum?

ii. [2 pts] What is the expected MS level 4 weeks after baseline for a participant who received the placebo gum?

(e) The linear predictor that you specified in 1(a) estimates a linear treatment effect over time for the placebo and active treatment groups. Since there were three distinct follow-up times, we could instead choose to model the effect of time as a factor.

i. [1 pt] Construct (and write down) a linear predictor which will allow you to assess whether the effect of treatment changes over time. In contrast to 1(a), model the time covariate as a categorical variable.

ii. Using the linear predictor given in (e) i, state the $\beta$ coefficients (or term involving the $\beta$ coefficients) which corresponds to the following hypothesis tests. (i.e. which $\beta$ coefficients, or terms involving the $\beta$ coefficients, equaling zero corresponds to the following).

A. [3 pts] No difference between the treatment gum and placebo gum groups in mean MS levels at 1 week, 4 weeks, and 12 weeks post-baseline.

B. [2 pts] No difference between the treatment gum and placebo gum groups in mean change from baseline to 4 weeks and mean change from 4 to 12 weeks post-baseline.

2. **Interpretation of GEE results [12 pts].** In this question, you will consider results from a study examining the informed consent process. In particular, this randomized trial studied the impact of a new informed consent process on the participants' understanding of study concepts in the setting of a preventive HIV vaccine trial. Participants were randomized to undergo the mock informed consent process or not. (Note that participants were not actually participating in an HIV vaccine trial, thus it was acceptable for some participants to not undergo the mock informed consent.) At study visits occurring at baseline and 6, 12, and 18 months post-baseline, participants were given a knowledge quiz to assess their understanding of study concepts. Our outcome of interest is whether participants responded correctly to a particular question, which asked them whether it was true or false that "The study nurse will decide who gets the real vaccine and who gets the placebo." This analysis focuses on the subset of 1123 participants who were injection drug users and thus considered at high risk of HIV infection.

   A GEE model was fit to estimate the effect of the informed consent process on the odds of answering the "nurse" question correctly. In particular, the following mean model was used:

$$\text{logit}P(Y_{ij} \mid \text{Post}_{ij}, \text{ICgroup}_{ij}) = \beta_0 + \beta_1 \text{Post}_{ij} + \beta_2 \text{ICgroup}_{ij} + \beta_3 \text{Post}_{ij} \times \text{ICgroup}_{ij},$$

   where $Y_{ij} = 1$ indicates that subject $i$ answered the question correctly at visit $j$, $\text{Post}_{ij}$ is an indicator for whether the visit was post-baseline (1=visits 2, 3, or 4 and 0=visit 1), and $\text{ICgroup}_{ij}$ indicates the study arm randomization (1=mock informed consent group and 0=control group). Note that $j = 1$ indicates the baseline visit (prior to the intervention), $j = 2$ indicates the 6-month visit, $j = 3$ indicates the 12-month visit, and $j = 4$ indicates the 18-month visit. Figure 1 presents the output from fitting this model.

   (a) [1 pt] What working correlation was used to fit the model whose results are provide in Figure 1?

   (b) [2 pts] Based on the results in Figure 1, calculate an estimate of the *probability* of a correct question response post baseline among control subjects and among intervention subjects. (Hint: expit(logit(a))=a where expit(a)=$\frac{exp(a)}{1+exp(a)}$.)

   (c) [2 pts] Using the results in Figure 1, provide an interpretation of $\hat{\beta}_1$ that is suitable for inclusion in a scientific publication. If necessary, transform the results to the most suitable scale for interpretation.

   (d) [2 pts] Provide an estimate with 95% confidence interval for the odds ratio comparing the odds of a correct question response after randomization relative to the odds of a correct question response at baseline among the intervention group. If you do not have sufficient information to calculate either of these, please indicate that and state what further information is needed.

   (e) [3 pts] What is $\exp(\beta_3)$ estimating? (Hint: Considering your answers to 2(c) and 2(d) may be helpful.)

   (f) [2 pts] Having not taken PubH 7430, your co-investigator instead fits the mean model specified above but using a logistic regression that does not account for the correlation between responses measured on the same participant. How would you expect the standard errors for $\beta_1$ and $\beta_2$ in the GLM model to compare to those of the GEE model?

Figure 1: GEE analysis results for informed consent study

Working Correlation Matrix

| | Col1 | Col2 | Col3 | Col4 |
|---|---|---|---|---|
| Row1 | 1.0000 | 0.2044 | 0.1936 | 0.1625 |
| Row2 | 0.2044 | 1.0000 | 0.3022 | 0.2755 |
| Row3 | 0.1936 | 0.3022 | 1.0000 | 0.3511 |
| Row4 | 0.1625 | 0.2755 | 0.3511 | 1.0000 |

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

| Parameter | Estimate | Standard Error | 95% Confidence Limits | | Z | Pr > \|Z\| |
|---|---|---|---|---|---|---|
| Intercept | 0.1676 | 0.0652 | 0.0398 | 0.2954 | 2.57 | 0.0102 |
| Post | -0.3238 | 0.0704 | -0.4618 | -0.1857 | -4.60 | <.0001 |
| ICgroup | -0.1599 | 0.1643 | -0.4819 | 0.1622 | -0.97 | 0.3306 |
| ICgroup*Post | 1.0073 | 0.2012 | 0.6128 | 1.4017 | 5.01 | <.0001 |