BIOS6643 **Final Quiz Solutions** **2016**

You are allowed to use any course material in completing this quiz. Do not contact other students to discuss the exam, even for clarification. If you have any questions regarding this quiz, please contact me at 303-398-1862, or send me an e-mail: strandm@njhealth.org. Regarding question #3, I have also attached the GzLMM likelihood detail notes, in Word version so that you could use it as a template and not have to build all of the equations from scratch (if you use Word). If you want to use something else like LaTeX, that is fine, but are more on your own. You can space out the questions below according to your needs. If you write your answers and will not be using Word or LaTeX, just use separate pages and clearly write the question numbers with your work. Unless other arrangements are made, your work is due back by 1pm. Have fun! Try to write sufficient, succinct answers.

(1) You conduct a longitudinal study involving subjects with and without a certain gene. It turns out that subjects with the gene drop out of the study more often than those without. In addition, the relationship between the outcome and time differs between those with and without the gene. If we have the genetic information for the subjects, do you expect there to be problems in estimating coefficients of interest in the longitudinal model (i.e., time-related coefficients) with respect to gene differences? Explain, and discuss what terms you would include in the model in order to carry out the analysis.

The data described above is missing at random (MAR). Basically, if we have the 'right' variables, we can include them in the model and avoid having the missing data cause bias in estimates of interest. I did not specify but let's say that simple linear relationships exist between y and time (x) for each gene group (i.e., we have x, group and x*group as predictors). Since we are essentially obtaining estimates that are group specific, it doesn't really matter than one group drops out more than the other. Note that if we did not include all of these terms in the model, then bias may be an issue. In addition, if the missingness does depend on responses not observed and there are no other variables to account for it, then we would have MNAR data; but just going by what is described, we have MAR data. 5 points

(2) You have binary longitudinal data, and you fit a model to estimate the relationship between the health outcome [y: currently sick (0), currently healthy (1)] and several key predictors, including nutrition level (x). You consider fitting the data using 2 different approaches: (i) one that includes a random intercept for subjects to account for general differences in health, and (ii) the other that does not have random effects but takes into account serial correlation of the repeated measures (e.g,. using an AR(1) structure).

a. Mention the model and method (real briefly) that you would use for these 2 approaches.

(i) Here we could use a basic GzLMM with a random intercept. Not only do we account for subject heterogeneity, but we also impose a covariance structure for the repeated measures.
(ii) Here you could use a GzLM with GEE, and employ the AR(1) working covariance structure.

5 points

b. For the approaches you suggested, can you compare model fits using goodness-of-fit statistics? Explain.

We cannot compare, for the 2 models I suggested. One will provide a -2 log likelihood and AIC (although approximated) and the other provides a QIC since it uses quaslikelihood estimation. In some special cases a GzLM can use regular likelihood estimation, but if you either add a dispersion parameter to a distribution that does not ordinally have it (e..g, Poisson), or you add the repeated measures with GEE (or both), we conduct quasilikelihood estimation. I didn't mention anything about a dispersion parameter, but we do know we have repeated measures. 5 points

c. It turns out that the variance for random intercept is relatively large. How would you expect this to affect the slope of *x*, and how would you interpret the slope for these 2 approaches (subject-specific or population-averaged)?

So I would expect it to attenuate the slope for the GEE model (approach ii) but would not expect to change the slope for approach i, since we have a random term to account for such differences. The estimate via GEE has a population-averaged interpretation, while the estimate for the GzLMM has a subject-specific interpretation. 5 points

d. Now say that you want account for both subject heterogeneity as well as serial correlation. What sort of approach would you take here?

Here I would suggest pseudo-likelihood estimation in the GzLMM, because we can then employ anything a linear mixed model method has to offer. In SAS, you would include a random statement for the intercept, and then a 'random residual' statement to account for the serial correlation. The default estimation approach in SAS PROC GLIMMIX is actually RSPL (the R indicating 'REML' type methods and S indicating a 'subject-specific' approach to estimation); i.e., if you do not include the METHOD option in the PROC GLIMMIX statement, that is what you will get. 5 points

(3) Consider an analysis to be done on longitudinal count data using Poisson regression in a generalized linear mixed model (GzLMM); a random intercept will be included for subjects and a simple linear regression for time will be included (including intercept, β_0 and slope, β_1). The probability mass function of the Poisson is $P(Y=y)=\lambda^y e^{-\lambda}/y!$ for y=0,1,2,…, where λ is mean and variance of Y. Write the likelihood function for the GzLMM in terms of β_0 and β_1. You will need to leave the integrands in your solution; as with the example given in class, use i to denote subjects and j for time.

$$
\begin{aligned}
L(\beta,\sigma^2;\mathbf{y}) = f(\mathbf{y}) &= \int l(\mathbf{y}\mid\mathbf{b})\mathrm{h}(\mathbf{b})d\mathbf{b} \\
&= \int_{\mathbf{b}} P(\mathbf{Y}=\mathbf{y}\mid\mathbf{b})\mathrm{h}(\mathbf{b})d\mathbf{b} \\
&= \int_{\mathbf{b}} \prod_{i,j} P(Y_{ij}=y_{ij}\mid\mathbf{b})\mathrm{h}(\mathbf{b})d\mathbf{b} \\
&= \prod_i \int_{b_i} \prod_j P(Y_{ij}=y_{ij}\mid b_i)\mathrm{h}(b_i)db_i \\
&= \prod_i \int_{b_i} \prod_j P(Y_{ij}=y_{ij}\mid b_i)\mathrm{h}(b_i)db_i \\
&= \prod_i \int_{b_i} \prod_j \left[\frac{e^{y_{ij}(\beta_0+\beta_1 x_{ij}+b_i)}e^{-e^{(\beta_0+\beta_1 x_{ij}+b_i)}}}{y_{ij}!}\right]\times e^{-b_i^2/2\sigma^2}/(2\pi\sigma^2)^{1/2}db_i
\end{aligned}
$$

Note that one of the key steps is identifying λ in terms of the Beta's and the random intercept (in our case it is the mean of Y, given b_i). The function needs to be approximated using quadrature or Laplace methods. Note that it is really the first and last lines above that I need; you don't have to include every line above for full credit. Regarding the 'y' exponent, it drops down since $[\exp(c)]^d=\exp(cd)$

The above is sufficient for full credit, but you can keep simplifying…

$$= \prod_i \int_{b_i} \prod_j \left[\frac{e^{y_{ij}(\beta_0+\beta_1 x_{ij}+b_i)-e^{(\beta_0+\beta_1 x_{ij}+b_i)}} e^{-b_i^2/2\sigma^2}}{y_{ij}!\sigma\sqrt{2\pi}} \right] db_i$$

$$= \prod_i \int_{b_i} \frac{e^{-b_i^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \left[\frac{e^{\sum_j y_{ij}(\beta_0+\beta_1 x_{ij}+b_i)-e^{(\beta_0+\beta_1 x_{ij}+b_i)}}}{\prod_j (y_{ij}!)} \right] db_i$$

10 points

(4) Regarding the previous question, suppose that we find the Poisson distribution to be too stringent for the model (specifically, the requirement that the mean and the variance are equal). Mention one alterative way to model longitudinal count data that offers more flexibility in fitting the data than the true Poisson.

There are many options here. One would be to add a dispersion parameter, for which we use quasi-likelihood estimation. Another approach would be to try the negative binomial distribution, which offers a bit more flexibility. 5 points

(5) A study is conducted on subjects that come in for multiple hospital visits. Ideally they should come in once a year, but it turns out that they often come in at different times or even miss visits. A longitudinal model will be used to examine health outcomes as a function of time and other predictors.

a. If the outcome can be modeled with a mixed model, what type of covariance structure would you suggest to account for repeated visits for subjects?

I would suggest using the spatial power structure or some sort of spatial structure to account for the unequal spacing. You could also random effects to the model, if it helps or is desired. 5 points

b. If we're now considering a binary outcome, how would you account for the repeated measures?

One way would be to use GzLMM/GEE, where we employ the AR(1) working covariance structure, but fill in the data with missing values so that records account for consecutive, equally spaced time points. Another approach would be to use pseudo-likehood estimation in a GzLMM so that spatial structures can be employed. 5 points