47/50 Good.

BIOS6643 **Final Quiz** **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.

Given that we have information regarding gene and time, and that we already know that there is an interaction effect between time and gene on outcome (i.e. effect of time on outcome differs by status of gene) we would be able to include the main effect terms (i.e. time and gene) and the interaction term (i.e. time*gene) within the model to determine the hypotheses of interest (perhaps include polynomial terms for time also). Because the design of the analysis itself is subject to selection bias, due to drop out from subjects with the gene, the analysis is subject to bias. The investigator needs to identify if the missingness is MNAR or not (i.e. MCAR or MAR). The trend regarding subjects with the gene and subjects without the gene should be investigated and the researcher should determine if there are systematic increases/decreases based on initial values and later values. If there are particular trends (i.e. change in slope) deemed systematic then the missingness is non-informative and the observed data will be everything that is needed to estimate the coefficients of interest, otherwise the missingness is informative then the information we have would not allow us to accurately estimate the coefficients. The two additional components that would be included in the model to assume a relationship between the slopes and the starting values are: 1) Use the starting value as covariate and 2) add random intercept and slope for time with covariance between the slope and time (type = UN) for G. Both approaches accounts for dependencies of the slope on the initial values.

3/5 Yes on the model; the description suggests MAR data.

(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)

Model:

$$Y_{ij}|b_i \sim Bernoulli(p_{ij}) for\ i = 1, \dots, n\ and\ j = 1, \dots, r$$

$$Y_{ij} = g(x'_{ij}\boldsymbol{\beta} + Z'_{ij}\boldsymbol{b})^{-1} + \varepsilon_i$$

$$\ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + \beta_2 time_i + b_{i0}$$

$$b_i \sim N(0, \sigma_b^2), \varepsilon_i \sim N(0, I\sigma^2)$$

Approach: Using a generalized linear mixed model framework (SAS: PROC GLIMMIX) with quadrature (The Whole Enchilada) method for approximating the likelihood.

ii) 5/5 Good, you remembered!

Model:

$$Y_{ij}|b_i \sim Bernoulli(p_{ij}) for\ i = 1, \dots, n\ and\ j = 1, \dots, r$$

$$Y_{ij} = g(x'_{ij}\boldsymbol{\beta})^{-1} + \varepsilon_i$$

$$\ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + \beta_2 time_i$$

$$\varepsilon_i \sim N(0, R_i); R_i\ with\ AR(1) structure.$$

Approach: Using a generalized estimating equation (GEE) framework (SAS: PROC GENMOD) with quasi-likelihood method for estimating the coefficients.

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

No. The quadrature (TWE) method uses numerical integration thus approximates the true likelihood, such as that of maximum likelihood method but the quasi-likelihood method does not build on maximum likelihood estimating techniques (i.e. computes estimates based on form of mean, variance as function of mean, working correlation parameters, and scale parameters) thus the computation of parameter and estimates will not be approximating the true likelihood thus the two models cannot be compared.

5/5

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)?

If the variance on the random effect is large, the population averaged function (averaged random effects) and the subject specific function (with random effects) estimates will not be the same. Thus, the slope of x would be different for population averaged and subject-specific models. The larger the variability the more stretched out the population averaged slope. 5/5

For approach i) because we are including subject-specific effects, the interpretation of the slope will be subject-specific given that we use a conditional model (including random effects). If we use a marginal

model, the slope parameters will have population-averaged interpretations. For approach ii), there are no random effects and only the marginal model will be used thus the interpretation of the slope would be population-averaged interpretation.

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

In this case we would use a generalized linear mixed model framework (SAS PROC GLIMMIX) with pseudo-likelihood estimation method to have both random effects (i.e. intercept and/or slope) and AR(1) structure for R_i.

5/5

(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.

$$\lambda_{ij} = \mathrm{E}(Y_{ij}|x_{ij}, b_i) = \exp(x_{ij}^r \boldsymbol{\beta} + Z_{ij}^r \boldsymbol{b})$$

$$L(\beta_0, \beta_1, \sigma^2; \boldsymbol{y}) = \prod_i \int_{b_i} \prod_j P(Y_{ij} = y_{ij}|b_i) h(b_i) db_i$$

$$= \prod_i \int_{b_i} \prod_j \frac{\lambda_{ij}{}^{y_{ij}} e^{-\lambda_{ij}}}{y_{ij}!} h(b_i) db_i$$

$$= \prod_i \int_{b_i} \prod_j \frac{exp(b_{i0} + \beta_0 + \beta_1 x_{ij})^{y_{ij}} e^{-exp(b_{i0}+\beta_0+\beta_1 x_{ij})}}{y_{ij}!} h(b_i) db_i$$

$$= \prod_i \int_{b_i} \prod_j \frac{exp(b_{i0} + \beta_0 + \beta_1 x_{ij})^{y_{ij}} e^{-exp(b_{i0}+\beta_0+\beta_1 x_{ij})}}{y_{ij}!} * \exp(\frac{-\mathrm{b}_i^2}{2\sigma^2})/(2\pi\sigma^2)^{1/2} db_i$$

9/10 Just clarify the 'y' exponent. See solutions.

(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 several methods that allow for correction of overdispersion.

i) Using a negative binomial distribution that exhibits increase dispersion.
ii) Add a normal random error into the likelihood function of the Poisson distribution.
iii) Use mixture model (zero-inflated Poisson).

5/5

(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?

A spatial power (SAS PROC MIXED REPEATED /SP(POW)(time)) or spatial exponential (/SP(EXP)(time)) structure. Spatial power structure works well with unequal spacing of time thus would be appropriate in this case.

5/5

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

Since the outcome is binary then the framework would now change to a generalized linear mixed model framework using SAS PROC GLIMMIX with quadrature (TWE) method, MODEL statement including a logit link function, binary distribution, and a RANDOM statement with _RESIDUAL_ and type=SP(POW)(time) as option would be possible. Thus a spatial power structure could be utilized in this case also.

Additionally, GEE framework could be used but then there is no spatial covariance structure that could be employed, only the AR(1). In this case data would need to be filled in for the missing points (data restructuring). This is too tedious.

5/5 It's not too bad, really...