**BIOS6643 Fall 2018 HW 6**

(1) The data will be collected on asthmatic subjects on every weekday for one month. Thus, the time is unequally spaced. There are two outcomes: (i) medication use **counts** and (ii) **FEV1**. The medication use counts outcome is non-normal count data, which should have 0 as the lower bound and likely to be right-skewed. The FEV1 is a continuous outcome with lower bound 0, which is a diagnosis criterion of lung function.

a. To take the within subject over time correlation into account:

For the outcome medication use **counts**: PROC GENMOD (To use GEE) or PROC GLMMIX METHOD = RSPL (GzLMM using linearization methods). In the MODEL statement, including DIST = poisson.

For **FEV1**: PROC MIXED

b. For **counts**: (under PROC GENMOD) REPEATED subject = id / TYPE = AR(1) modelse; (model based s.e.). To deal with unequal spacing in GzLM/GEE, I need to include records for equally spaced time points and fill in with missing values for weekends.

For **FEV1**: repeated / type = sp(pow) subject=id;

c. In this case, the medication indicator is a binary outcome. If we are not interested in the time serial correlation, we can use PROC GLIMMIX METHOD = QUAD, MODEL / distribution = binary. Random intercept / subject = id; This approach is to use adaptive Gaussian quadrature to numerically approximate the log likelihood.

Drawbacks: You need to consider the best number of quadrature points to use. The more, the better, although as you increase it, the computational burden increases. This is arbitrary and computing intensive. By this approach, you only can specify the random terms, and you cannot specify the variance structure of the outcome (V) by the R matrix, no repeated statement.

d. To include both random terms and repeated serial correlation, you should use the PROC GLMMIX METHOD = RSPL (GzLMM using linearization methods); random intercept / subject = id; random \_residual\_ / subject = id type=sp(pow)(day);

Drawbacks: This is a doubly-iterated procedure, more complex. One drawback to the linearization method is estimator bias that has been reported. However, as sample size increases, the bias diminishes quickly.

(2)

a. Normal:

E(*Yij*| **x***ij*, **b**i=0) =

E(Yij|xij) =

b. Binomial

E(*Yij*| **x***ij*, **b**i=0) =

E(Yij|xij) =

c. Poisson

E(*Yij*| **x***ij*, **b**i=0) =

E(Yij|xij) =

(3)

a. For 10 units change, interpretation, exp(10\*0.0019) = exp(0.019) = 1.02, 2% increase in albuterol use for a 10 μg/m3 increase in mmaxpm25.

b. The IQRs for mmaxpm25, temperature and humidity are 11, 16 and 29, respectively. And mmaxpm25 0.0019, temp -0.0060, and humidity -0.0036. Here, I want to get the slope estimates per 20 standardized units change.

For mmaxpm25, exp(20\*0.0019/11) = 1.0035, 0.35% increase in albuterol use for a 20 standardized unitincrease in mmaxpm25.

For temp, exp(20\*(-0.006)/16) = 0.9925, 0.75% decrease in albuterol use for a 20 standardized unitincrease in temperature.

For humidity, exp(20\*(-0.0036)/29) = 0.9975, 0.25% decrease in albuterol use for a 20 standardized unitincrease in humidity.

c. Since the residual has a Poisson distribution, and a log link function. When there are only random intercept differences for subjects in the population, estimates of predictors have have both population-averaged and subject-specific interpretations. Because the only difference of means between population average and subject-specific average subject is in the intercept. No difference in the slope.

(4) Binary outcome, odds ratio.

a. The method here is GzLMM using pseudo-likelihood (the linearization approach), with a random intercept for subjects and a spatial power variance structure accounting for the time serial correlation. The estimates of day and weekend by SAS are, -0.00228 and -1.6514, respectively. And the estimates are the same by R. The results won’t change if the day effect were rescaled to a week or month. The R and SAS procedures match.

If we only include the random intercept, the estimates will differ. In this context, rescaling of the day effect to a week or month will attenuate the time serial correlation. Thus, we will have closer estimates.

exp(-0.00228\*10) = 0.977, thus 10 days later the odds of exacerbation decreases by 2.3%, for a given subject.

exp(-1.6514) = 0.192, thus 1 weekend later, the odds of exacerbation decreases by 80.8%, for a given subject.

b. For a binary outcome and logit link function, the slope estimates here have different interpretations. The marginal model has parameters with population-averaged interpretations, while the conditional model has subject-specific interpretations. And this is because the logit link function results that population average and subject-specific average subject have difference in both intercept and slope.

c. If fit the data with GzLM/GEE, the time serial correlation cannot be considered by R matrix. The beta estimates would be similar. However, due to the mis-specification of covariance structure, model-based SEs will be much bigger. This is probably due to the dispersion will not being modeled properly without R-side parameters in the model.

(5) Name one advantage and one disadvantage of fitting the exacerbation discussed above using a GzLMM with Gaussian quadrature. What estimation approach is used?

Advantage of GzLMM with Gaussian quadrature, using more quadrature points, one can get unbiased and accurate beta estimates.

Disadvantage: this approach only accounts for random effects, cannot include the spatial power variance-covariance structure.

The estimation approach here is to use numerical integration method, Gaussian Quadrature, to approximate the true likelihood function. An optimization technique such as the Dual Quasi-Newton method can then be used to maximize the approximate function in order to determine maximum likelihood parameter estimates. Finally, the random effects are estimated as the mean of the random-effects distribution, given the data and parameter estimates, just as the way of standard LMM method.