**BIOS6643 GLM (Page 1) and LMM (Page 2) Review Questions Fall 2018**

**NOTE: To turn in: 3, 8, 9, 11, 12 (a and b only), 14. The others are for practice and discussion.**

1. Consider using PROC GLM in SAS to fit a regression model. If one of the predictors is gender, coded as an indicator variable (e.g., ‘1’ for Female and ‘0’ for Male), you will essentially get the same model fit whether or not you put this variable into the CLASS statement. Thus, although it is clearly not a continuous variable, we can treat it as such when fitting the model. Briefly describe why this is the case.
2. Complete the practice question in the GLM chapter of the course notes on the bottom of page 63.
3. **The distri**bution of is given on page 62 of the GLM course notes, which was derived using the linear form result. Derive the distribution of  if  form some vector **a**. Note that the result can also be completed using the linear form result. The proof is short.
4. We’ve discussed how in some cases, the approach of simplifying a general linear model that has class variables up front yields the same result as using a g-inverse, e.g., setting the highest level(s) of factor(s) to 0 is essentially equivalent to the way SAS computes the g-inverse. In the set-to-0 approach, we know that estimates for levels other than the level that did not have an indicator reflect comparisons to that level (i.e., the one without the indicator is the ‘reference level’). Using estimability and the less-than-full-rank model, show how we know this to be true. [Hint: consider estimability of κi–κj in the one-way effects model.]
5. Consider a 3×2 factorial experiment with 2 replicates in each treatment combination. The data will be analyzed using the model , *i*=1,2,3; *j*=1,2; *k*=1,2 (*k* denotes the replicate).
   1. Write the **β** and **X** matrices for the general linear model for the *effects model* shown above.
   2. What is *r*(**X**)?
   3. Are the following estimable? Justify your response.
      1.  ii.  iii.  iv. 
   4. For the bread data (Neter, p. 686; posted on web page) determine the following using SAS PROC IML, R, or other software, based on the less-than-full-rank model written above.
      1. 
      2. S.E. ()
6. Complete the practice quiz at the end of the GLM chapter of the course notes, page 101.
7. Consider a study or experiment that has two factors (e.g., group and time); each factor has 4 levels and will be treated as a class variable. We will create a model for response ‘y’ as a function of *group*, *time* and *group\*time*; there are 3 replicates for each group-time combination.
   1. How many columns are in **X** for the less-than-full-rank model?
   2. If you were to write a full-rank statistical model for these data, how many parameters would there be? (I.e., how many columns are in **X** for a full-rank model?)
8. **For the Myostatin da**ta, note that the population mean for the myostatin group at 48 hours is  for the one-way effects model (see the course notes). Write the population means for the following. NOTE: these are not numerical estimates, but parameters or combinations of parameters.
   1. Myostatin group at 48 hours; means model.
   2. Myostatin group at 48 hours; two-way effects model.
   3. Myostatin group, difference between 48 and 72 hours, one-way effects model.
   4. Myostatin group, difference between 48 and 72 hours, two-way effects model.
9. **Show that**  satisfies the normal equations. (Here, tilde indicates that the beta estimate may not be unique.)
10. Top 5 race times by individual age and gender were recorded for the 1995 Bolder Boulder 10K race. We are going to model race time as a function of age and gender, including linear and quadratic terms for age, as well as age×gender and age2×gender interaction terms, for 30 to 60 year-old males and females. Here, we don’t have a random sample; these are extreme values since they are the fastest times for each age, but we are more concerned with curve fitting than inference.
    1. Write the model in terms of a single outcome (i.e., not in matrix form).
    2. Write **β** and the first 10 rows of **X** for the matrix model (get the data from the course web site).
    3. Analyze the data with PROC GLM. Write the fitted functions for 10K race times by age separately for men and women.
    4. Graph the data using either SAS or R. Use different symbols for males and females and superimpose the fitted functions.

*LMM models*

1. **Review Section 3.6.3** in the GLM course notes. (Note: this should be in the LMM chapter since it is discussing models indexed by subject and time, i.e., applicable for repeated measures.)
   1. Write full-rank and less-than-full-rank models if there is a group variable with 4 levels (i.e., 4 groups), a time variable that is treated as a continuous variable (linear term only), plus group\*time interaction. How many columns are in **X** for each approach?
   2. If time points are unequally spaced then would it be appropriate to treat time as a class variable? Explain.
2. **Consider a study where subjects** in 3 groups (e.g., race or treatment) are observed over 3 times and some health outcome, y, is measured. Unless otherwise mentioned, include a random intercept for subjects to account for the repeated measures. For simplicity, use 2 subjects per group.
   1. **Consider** modeling group and time as class variables, plus interaction. Write statistical models and the X matrix for the following cases.
      1. No restriction placed on the model. I.e., write the less-than-full-rank statistical model.
      2. A set-to-0 restriction is placed on the parameters associated with highest levels.
      3. A sum-to-0 restriction is placed on the parameters associated with highest levels.
3. **Show that** the linear trend for one group compared to another (say Group A versus B) is estimable by showing that **L**=**LH**, where the Moore-Penrose inverse is used in calculating **H**. First you need to construct **L**. (As a check, repeat using SAS’s g-inverse in calculating **H**, but you don’t need to turn that in.
4. Say that both Group and Time are treated as ‘continuous’ (i.e., not included in the CLASS statement in SAS or factor argument in R). How does this change the model and X matrices?
5. Write the statistical models in 12ai if an AR(1) structure for **R** is included.
6. Derive  in an LMM, given the algebraic form of  that is obtained via ML estimation. NOTE: there are two types of variance, model-based and empirical. The difference is whether the middle ‘V’ is determined via the model or using squared residual quantities; derive the model-based form. To answer this question, work with the ‘complete data’ form of .
7. **For the either the** Dog data or Beta Carotene data, design and compute 2 contrasts and 2 estimates (other than those done in class or previously). Create your tests and estimates based on what you think is interesting. With the output, write up your results in a few sentences.