MS Take-Home Comprehensive Exam (Statistics)

Due by NOON on 21 August 2017

Instructions:

Read each question carefully and ask for clarification as needed. If you have questions during the weekend, please send them in an e-mail addressed to both Jenny (jgreen@montana.edu) **AND** Laura (laura.hildreth@montana.edu) so that we can provide consistent responses. We will not respond to e-mails sent to us individually.

You are allowed to use any resources from Stat 501/502 and 505/506, the internet and resources from any other course. **If you do use non-Stat 501/502 and 505/506 resources they must be referenced, including internet resources. You may not receive help from other people except Jenny and Laura. Do NOT discuss this exam with other students until after noon on Monday 21 August 2017!**

When answering questions, please number your responses (you do not need to include the questions) and organize them in sequential order. Answers may be typed or written, but all work must be neat, organized and legible. When applicable, all work must be shown and all explanations (including assumptions, known distributional results, and theorems), plots and/or code must be provided in order to receive credit. (That is, be sure to show all work and justify/explain answers.) **When using R or SAS, please include your code in an Appendix, annotated and organized by question number.**

Please compile your numbered answers and Appendix in a single document and e-mail the file to John Borkowski (jobo@montana.edu) by noon on August 21, 2017. By turning in this exam, you acknowledge that you have completed this exam in accordance with the Student Conduct Code for Academic Honesty found online at http://www.montana.edu/policy/student_conduct/academicmisconduct. **Failure to comply with this code will result in an automatic score of 0, failure of the comprehensive exam, and you will be reported for academic dishonesty.**

****

1) **Diseased Pine Trees Revisited** [20 pts]. The in-class portion of the exam described the following situation: An ecologist studying diseased pine trees takes data (a_i, Y_i), $i = 1, 2, \ldots, n$, where a_i is the known size of the i^{th} randomly selected area and Y_i is the number of diseased pine trees in that area. Suppose we model the data by $Y_i \sim Poisson(a_i\theta)$, where the $Y_i s$ are assumed to be independent, and we want to estimate θ^2. Earlier, we proved $S = \sum_{i=1}^{n} Y_i \sim Poisson\left(\theta \sum_{i=1}^{n} a_i\right)$, and we are now considering the following two estimators:

   - the maximum likelihood estimator (MLE) of θ^2: $T_1 = \dfrac{S^2}{\left(\sum_{i=1}^{n} a_i\right)^2}$, and
   - the uniformly minimum variance unbiased estimator (UMVUE) of θ^2: $T_2 = \dfrac{S^2 - S}{\left(\sum_{i=1}^{n} a_i\right)^2}$.

   Which estimator (MLE or UMVUE) of θ^2 would you recommend using? Why? Use simulation and theory to support your answer with appropriate statistical evidence. Show and justify all work, being sure to explain how you obtained your simulated results (and what they represent) and why the selected criteria/evidence indicates one estimator is "better" than the other. **Please remember to include your code in an Appendix.**

2) **GCSE** [40pts]. Goldstein et al. (1993) conducted a study using data from 4059 students from 65 different schools in inner London. In the UK, as well as several British territories, students take an examination over several subjects in order to earn a General Certificate of Secondary Education (GCSE). Of interest in this analysis is the total score (0-70 points) received for all subjects taken during the examination. These scores were then standardized to have a mean of 0 and a standard deviation of 1 (in other words a z-score was calculated for each student using the sample mean and sample standard deviation to standardize). Let Y_{ij} be the standardized total score for the i^{th} of n_j high school students $\left(i=1,2,...,n_j\right)$ in the j^{th} of k schools $\left(j=1,2,...,k\right)$. We will consider two (non-random) covariates: x_{1ij} is the student's standardized score on the London Reading Test (LRT) which is a common reading test taken by students when they are 11 years old, and x_{2ij} is the gender of the student (1=male, 0=female). The data are in the file exam.csv. **As before please remember to include your code in an Appendix.**

To analyze these data, let's initially consider the multilevel or hierarchical linear model:

$$Y_{ij} = \left(\beta_0 + u_{0j}\right) + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \varepsilon_{ij}$$

where β_0, β_1, and β_2 are unknown parameters and $\varepsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma_e^2\right)$.

a) In this model, students are assumed to be nested within schools. Explain what this means in context of the problem and why this is a reasonable assumption. [4pts]

b) Explain what the random effect u_{0j} represents and what assumptions we (typically) make regarding this random effect. Briefly explain why a researcher would include this effect in the model and why it is assumed to be a random effect as opposed to a fixed effect. [6pts]

c) Derive $Cov\left(Y_{ij}, Y_{i'j}\right)$, $1 \le i < i' \le n_j$. Show all work, and interpret what this represents in context of the problem. [6pts]

d) Using the data, provide an estimate for the parameter of interest in part c. [2pts]

e) When obtaining a point estimate for part d you (should have) used REML estimation. Carefully explain how this estimation method differs from maximum likelihood and why it is typically used. [5pts]

f) Using the data, explain whether between or within school variability is the larger source of variability in standardized exam scores. Support your answer with appropriate evidence. [6pts]

g) Use large sample normal distribution theory (Wald method) to derive a 95% confidence interval for β_1. Show and explain all work, being sure to state any assumptions and theorems used. [8pts]

h) Using the interval derived in part f and the data, calculate and interpret a 95% confidence for β_1. [3pts]

3) **GCSE Revisited** [70pts]. Another option to use to model the data described in question 2 is the multilevel or hierarchical linear model:

$$Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})x_{1ij} + \beta_2 x_{2ij} + \varepsilon_{ij}$$

For the remainder of this exam use this model! Don't forget to provide your code in an Appendix.

a) Explain how this model differs from the model initially considered. That is, what does u_{1j} represent, what additional assumptions do we make when including this effect, and why we want to consider this effect random as opposed to fixed? [6pts]

b) We would like to test if it is necessary to include u_{1j} in the model or equivalently we would like to conduct a test to compare the initial model to the current model.
   i. Create a plot (or set of plots) to visualize whether the data suggest that we should include u_{1j} in the model. Based on the plot(s), do you think this additional effect should be included? Describe which aspects of the plot(s) support the inclusion (or exclusion) of u_{1j} and/or u_{0j}. [8pts]
   ii. Set up the hypotheses we want to test in context of the parameter(s) of interest. That is, do not write out the hypotheses in terms of the models being compared but rather in terms of parameters. [2pts]
   iii. To conduct this test it is not recommended to use a standard likelihood ratio test. Why? [4pts]
   iv. Provide an alternative way that we could compare these two models. Using this alternative approach, explain whether it is necessary to include u_{1j} in the model. [6pts]

c) Find the exact distribution of Y_{ij}. Show and explain all work, being sure to state any assumptions and theorems used. [8pts]

d) Derive $Cov(Y_{ij}, Y_{i'j}),\ 1 \le i < i' \le n_j$. Explain why this result differs from the one obtained in question 2, part c. When deriving this covariance, show all work and state any assumptions used. [10pts]

e) Harriet (a girl) attends a school in inner London and is studying to obtain her GCSE. Her (standardized) LRT score was a 1.37. What standardized total score should Harriet expect to obtain on her exam? [3pts]

f) Harriet not only attends a school in inner London but attends school 8 in this study. Using this information what standardized total score should Harriet expect to obtain on her exam? [3pts]

g) Carefully explain why your results from parts e and f are not the same. [5pts]

h) One potential concern is that the variability of standardized total scores may differ across schools (e.g., school 3 may have more variability in its scores than school 42). Propose a way that the researchers could incorporate this concern into their model. You do NOT need to implement what you propose. [5pts]

i) Prior research indicates that students' performance in mathematics is often related to their performance in other subjects, so the researchers are interested in accounting for the potential effect of a student's mathematics teacher (i.e., the teacher primarily responsible for the student's mathematics instruction). Advise the researchers on how to proceed with such an analysis by addressing the following questions:
   i. What additional information (or data) are needed? [2pts]
   ii. When incorporating teacher effects into the model, should they specify them as random or fixed effects? Why? How should these effects be included in the model? (Note that you don't have to write out the model equation-just describe in words how it should be included in the model.) [5pts]
   iii. What are possible benefits and/or disadvantages to incorporating teacher effects into the model? [3pts]