Methodology Minor Field Exam

Fall 2018

For the minor field exam, you must answer two questions, one in the morning session and one in the afternoon session. In the afternoon session, you may use the software of your choice. You are free to use whatever word processing or typesetting software you like to write your answers. The questions must be answered in the allotted time. Please give your student ID (810) number, not your name, on the exam.

For the morning session, internet usage is prohibited. For the afternoon session, you may use the internet to download software packages or look up reference information as you complete the data analysis. Your work must be fully your own. Enjoy this opportunity to showcase your skills!

Morning Session: Statistical Theory and Modeling Decisions

Answer **one** of the following two questions. Please provide the **number** of the question you are answering. Reproducing the question text is not necessary.

1. *Bayesian Statistics:* An essential component of Bayesian inference is that the researcher must formally specify priors for all model parameters. In some circles, priors are viewed as problematic while, in others, they are viewed as a feature. What are some of the main objections to the use of priors and why are some people opposed to them? What are some reasons that people argue priors are beneficial for theory testing/inference? Give one example of when one might use a prior for reasons other than inserting researcher judgment.

   Describe three approaches by which a researcher could formulate priors for the parameters of a model? Explain the process of constructing a prior under each of these three approaches and discuss their relative strengths and weaknesses.

   Now describe a dataset/model that you would want to model using Bayesian statistics. What parameters would require you to specify priors and how would you approach developing these priors under the three approaches you described above? Which approach do you feel is best and why?

2. *Linear Model Theory:* When estimating a linear model via OLS, one assumes the Gauss-Markov assumptions to be true in order for the estimates to be BLUE. What are these assumptions and what does it mean for an estimator to be BLUE? What

are the most common violations of Gauss-Markov and what types of data are most likely to lead to violations of these assumptions? What are the implications to violating Gauss-Markov, specifically in terms of interpreting model results? De- scribe at least two tests for determining whether Gauss-Markov holds for a given model/dataset. Finally, which assumption do you feel is the most important to not violate and why?

Afternoon Session: Analyzing Data

Answer **one** of the following two questions. Please provide the **number** of the question you are answering. Reproducing the question text is not necessary.

1. Please analyze the **Salary** data set using a linear model estimated with ordinary least squares. The outcome of interest is the variable *salary*, which is the annual salary of small university professors measured in dollars. The input variables are:

   - sex–coded 1 for female
   - rank–1 for assistant prof, 2 for associate and 3 for full
   - years in rank–number of years in current rank
   - degree–highest degree earned coded 1 for PhD and 0 for MA
   - years since degree–number of years since highest degree was earned

   You must specify the model using all 5 input variables. Additionally, test the hypothesis that the effect of *rank* is conditional on the *sex* of the professor and provide a discussion of this test. Present the results of this model in a publication quality table including the coefficients, standard errors, R^2, at least one other measure of model fit, and any additional information you would like. What is the substantive meaning of each coefficient? What can you conclude from the *t*-ratios associated with each coefficient? What does the R^2 tell you? Also, please assess the degree to which multicollinearity, heteroscedasticity, and endogeneity pose threats to inference. Also assess whether or not there are any influential cases/outliers that could potentially be biasing the results. Be sure to explain how you diagnose these potential problems (be it graphically or with a hypothesis test) and report the results of your diagnoses.

2. Please analyze the **Admit** data set using a logistic regression model. This data set includes information about whether or not students were accepted to graduate school. The outcome of interest is the variable *admit*, coded 1 for admitted and 0 for not admitted, and the input variables are:

   - gre–the student's gre verbal test score
   - gpa–the student's graduating GPA

- rank–a categorical variable measuring the prestige of the student's college. Category 1 indicates high-prestige institutions with prestige declining as the levels increase. Category 4, therefore, represents the lowest-prestige institutions

You must specify the model with all 3 explanatory variables, being sure to properly take account of the level of measurement. Present the results of this model in a publication quality table including the coefficients, the standard errors, at least one fit statistic, and any additional information you would like. What can you conclude from the *z*-ratios associated with each coefficient? Please assess the effect of each input variable using odds ratios or predicted probabilities using tabular and graphical displays. Also, explain why your preferred interpretation method is the most informative.