Methodology Minor Field Exam

Spring 2019

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.

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:

1. *Time Series:* With multivariate time series analysis, two approaches are common: structural equation models and vector autoregression. Describe the procedure to each approach, and be sure to address these points: How does each approach view the role of theory in model specification? Relatedly, how are models specified? Once the model is specified, what is an estimator that each approach might use to find parameter values?

   Suppose that two scholars, Jim and Patrick, are examining monthly data on the environment and public health in Tbilisi, Republic of Georgia. The three variables are COPD rates (or lung disease), the levels of lead in the air, and the average temperature. Both Jim and Patrick agree that these variables should be related, but they take different approaches. Jim uses a structural equation model for these data, and Patrick uses a vector autoregression.

   How do you think Jim would have specified and estimated his model? How would Patrick have specified and estimated his model? What are the strengths of each approach? If the President of Georgia wanted a report on how these variables, which of the two models would you present to her and why?

2. *Bayesian Statistics:* After estimating a Bayesian model with MCMC, we normally assess convergence. Why is it important to do this? That is, what is the statistical reason why this concept is important?

   Name three means of evaluating convergence. (Note that each of these can be either a significance test or a visual examination of some sort.) For each of these three tests, please answer the following questions: What do you need in order to conduct the test, in terms of chains and number of iterations? When conducting the test, what is the intuition behind how the test or examination works? (You need not report a formula, just the intuitive idea.) When, in general, do you conclude that the model has not converged? When are you satisfied that it has converged?

   Suppose you find evidence of non-convergence on a test. What would you do next? What do you do when you are satisfied that there is convergence?

Afternoon Session: Analyzing Data

Answer one of the following two questions:

3. *Linear Regression:* Please analyze the data set *discrim.dta* using a linear regression model. The variable of interest is the variable *SAL77*, the 1977 annual salary of 32 male and 61 female employees at the Harris Bank of Chicago measured in dollars.

   The input variables are as follows (you must use them all):

   FSEX : Gender of each employee (1 for females, 0 for males).

   EDUC : Education level of each employee, measured as years of education.

   SENIOR : Seniority of each employee measured in months since first hired.

   AGE : Age of each employee measured in months.

   Present the results of this model in a table including the coefficients, the standard errors, the R^2, and any additional information you would like. What can you conclude from the t-ratios associated with each coefficient? What can you conclude from the model fit?

   Next, please test the conditional hypothesis that the education level of employees influences salary, but this effect is conditioned by whether the employee is male or female. Estimate a new model to test this hypothesis and discuss the results. Illustrate the nature of this conditioned relationship by graphing predicted values and confidence intervals. Provide a detailed interpretation of the conditional relationship and whether or not you think it matters.

   Then compare the fit of the two models and discuss the implications of including the conditional relationship described above relative to not including this. Which model do you feel is a better fit to the data and why?

   Finally, discuss whether or not you think OLS is the appropriate estimator for these data. If so, justify your response. If not, what model do you think would be a better estimator and why?

4. *Count Model:* Please analyze the data set *unrest.dta* using a count model. The outcome of interest is the variable *unrest*, a count of protest events in a given country.

   The input variables (you must use them all) are:

   CL : Freedom House civil liberties index (1 - 7 scale, with higher values indicating lower levels of civil liberties).

   soviet : Dummy variable indicating whether a country is a former Soviet block country

   polity : an index that ranges from -10 to 10 measuring level of democracy (higher values = more democratic)

   politysq : polity squared

   urbanpop : Percentage of a country's population that lives in an urban setting

   Start by fitting a Poisson model and reporting these results. Please test for overdispersion in these data. Describe what overdispersion is and why it is potentially a problem. What conclusions can you draw from these tests? What is the best choice of count model for these data and how did you make this choice?

   For every set of results you report, present the results in a table (separate or combined, across models) including the coefficients, the standard errors, at least one fit statistic, and any additional information you would like. For the one model you determine to be best for these data, please tell us: What can you conclude from the z-ratios associated with each coefficient? For all models, what can you determine from the fit statistic?

   Now test the hypothesis that the effect of Civil Liberties on unrest events is different in former Soviet countries than in the rest of the world. Test this hypothesis using the count model that you determine to be the best for these data. Please illustrate the nature of this conditioned relationship using predicted counts with confidence intervals. For this one model, please assess the substantive effect of all the other input variables as well. When interpreting the effects of other predictors, you may choose among the methods of: partial changes in the conditional mean, factor change in the conditional mean, discrete change in the conditional mean (e.g., predicted counts), or predicted probabilities of counts.