Methodology Minor Field Exam

Spring 2015

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 and will have access to the internet–which you may use to help you analyze data but NOT to communicate with anyone. You are free to use whatever word processing software you like to write your answers. The questions must be answered in the alloted time.

Morning Session: Statistical Theory and Modeling Decisions

Answer one of the following two questions:

1. *Bayesian Statistics:* Bayesian approaches to statistics have become increasingly popular in recent years. The Bayesian framework, while it is often used to answer the same questions to which frequentist methods have been applied, rests on fundamentally different philosophical foundations and uses different methods for estimation. In your view, which of these two things (the foundations or the estimation methods) is the most attractive feature of the Bayesian approach? Be sure to discuss the main differences in both of these things, commenting on both the advantages and disadvantages of the Bayesian approach relative to the standard frequentist approach in terms of fundamental differences as well as estimation methods and applicability to modern problems in political methodology.

   Next, consider a linear model in the Bayesian context where income is the dependent variable and gender, race, and education are the independent variables. Describe, in detail, how a Gibbs sampler would be used to estimate the coefficients in this model. Additionally, write out all necessary steps to estimate this model. Once estimated, describe the process necessary to make valid inferences from these results. Specifically, what are the main differences in interpretation and model checking between the frequentist and Bayesian approaches? If there are additional steps required in the Bayesian estimation, carefully describe what these are.

2. *Causal Inference:* Experiments are sometimes referred to as the "gold standard" for establishing causal links. In an ideal experiment, researchers can manipulate a single independent variable and assess its affect on a dependent variable without worrying about anything affecting the dependent variable. Unfortunately, true experiments are often not possible in the social sciences. Statistics do allow researchers to attempt to establish causal links.

   a. In what ways do generalized linear models attempt to mimic experiments? In what ways do linear models fail to clearly establish causation.

b. Using linear models to establish causality can especially difficult in the face of reciprocal causation, or in other words, a situation where an independent variable may affect the dependent variable but the dependent variable may also affect the independent variable. Choose an area of study of interest to you where reciprocal causation is a concern. For instance, many scholars believe that nations economic development leads to the development of democratic institutions, and nations expansion of democratic institutions leads to economic development. For your example, outline the problem of possible reciprocal causation and discuss how this makes causal inference problematic.

c. There are a variety of methods designed to deal with the problem of reciprocal causation, but each works best with different varieties of data. Name three methods for conducting causal inference when addressing this problem of reciprocal causation. For each method please answer three subquestions: For what kind of data is the method most suitable? How is the method implemented in practice? What are the strengths and weaknesses of the method?

d. Consider the example you consider in part b. Which method of dealing with reciprocal causation would be most appropriate for this example. Describe the data from your example and explain which estimation method would be the best to use with this data. Do you still have concerns about establishing causality? Why or why not?

Afternoon Session: Analyzing Data

Answer one of the following two questions:

3. Please analyze the data set *cabinet duration.dta* using a linear regression model. The data set contains information on how long governments survived in office for 15 countries in the post-war period. The variables are as follows (you must use them all):

   - durat: Duration of government in months (dependent variable)
   - polar: Polarization index [0-5]: a measure of support for extremist parties
   - fract: Fractionalizatio [0-5]: an index characterizing the number and size of parties in parliament such that higher values indicate dispersion into a larger number of relatively smaller blocks.
   - format: Number of attempts required to form a government
   - minority: Majority/minority government (1=Majority, 0= Minority)

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

   It has been suggested that minority governments tend to last longer when there is a lot of fractionalization. Estimate a new model to test this hypothesis and discuss the results. Illustrate the nature of this conditioned relationship by graphing predicted values and confidnece intervals. Provide a detailed interpretation of the conditional relationship and whether or not you think it matters

   Next 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?

   Next, assess whether or not there are problems with collinearity and heteroskedasticity. Also, check for outliers/influential data points–you may refer to the data points by row number as you do not know which country/year each row represents. Include the appropriate graphs/tables and be sure to discuss the results of these tests in detail.

   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. 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, and 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–whether a country is a former Soviet block country
- polity: an idex 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 and 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. For the one model 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 following 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.