Methodology Minor Field Exam

Spring 2016

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:* An essential component of Bayesian inference is that the research must formally specify priors for all model parameters. While some scholars of the frequentist or likelihoodist ilk may object to the insertion of researcher judgment at this stage, many Bayesians would argue that priors are a strength of the Bayesian approach. What are some of the main objections to the use of priors? What are a couple of reasons why priors can be beneficial for 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 Bayesian model. Explain the process of constructing the prior. What are the relative strengths and weaknesses of the three approaches?

   Describe a real or imagined dataset that you might want to model using Bayesian statistics. Your example data can have any kind of features you think would be illustrative (e.g., pure cross-section, time series, panel data, geospatial data, multilevel data, etc.). Suppose you were specifying priors for your example data: Describe what you would do specifically if you followed one of the three approaches described earlier. Why is the approach you are choosing the best option?

2. *Measurement Theory:* Clyde Coombs (1964) considered data as relations between points in space. Based on this premise, he classified relational data into four kinds, based on two criteria. The first criterion is whether the observed data consist of observed points from two sets (individuals and stimuli, such as legislators and bills) or one set (only stimuli). The second criterion is whether the data consist of pairs of distances between points (such as ideological distance) or pairs of points themselves (such as right or wrong answers on a test). The four kinds of data are thus:

   I Preferential choice data, or individual-stimulus differences comparison. (Pairs of distances, with points from two sets.)

   II Single simulus data, or individual-stimulus comparison data. (Pairs of points, with points from two sets.)

   III Stimulus comparison data. (Pairs of points, with points from one set.)

   IV Distance comparison. (Pairs of distances, with points from one set.)

   Consider the first three types of data (preferential choice data, single stimulus data, and stimulus comparison data). Choose two of the three types of data and answer all of the following questions for each of the types you consider:

   - Describe a real or hypothetical dataset that fits the description of this type of data. Explain how your data's features correspond to each of the two criteria.
   - What is an appropriate measurement method to apply to your example data? Describe why the method is appropriate for this type of data. What are the steps of your proposed method?
   - Explain what types of information these methods will yield and how one would present and interpret these results.

Afternoon Session: Analyzing Data

Answer one of the following two questions:

3. Please analyze the data set *incumbent.dta* using a logistic regression model.The data are available here:

   `http://spia.uga.edu/faculty_pages/monogan/teaching/incumbent_logit.dta`

   The data set contains information on members of the House of Representatives in 1990. The outcome of interest is the variable *returned*–whether the member was returned to the House in 1992, and the input variables are (you must use them all):

   - age: The incumbent's age
   - surplus: Surplus the incumbent could take home if they retired
   - marginal: Dummy-whether the race was close last time
   - resistm: Dummy-whether the incumbent's district was redistricted

   Present the results of this model in a table including the coefficients, the standard errors, the proportional reduction in error, 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 proportional reduction in error?

   Please test the conditional hypothesis that the effect of redistricting varies according to an incumbent's age. Please illustrate the nature of this conditioned relationship using predicted probabilities with confidence intervals. Please assess the substantive effect of the other input variables as well, reporting odds ratios and predicted probabilities. What are the tradeoffs of these two interpretation techniques?

   Next, estimate the same model but *without including the interaction term* and use a likelihood ratio test to determine whether the more complex model offers a significantly better fit. Lastly, if you wanted to do some kind of residual analysis, how would you go about that in principle?

4. Please analyze the data set *unrest.dta* using a count model. The data are available here:

   `http://spia.uga.edu/faculty_pages/monogan/teaching/unrest_count.dta`

   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 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.