Methodology Minor Field Exam

Fall 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. 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.

Morning Session: Statistical Theory and Modeling Decisions

Answer one of the following two questions:

1. *Bayesian Statistics:* Hierarchical modeling has become increasingly popular in recent years, and Bayesian hierarchical modeling has risen in popularity as part of this trend. Within Bayesian hierarchical models, we can create a substantive hierarchy, a methodological hierarchy, or both. What do we mean by each kind of hierarchy? Describe an example of a model with a substantive hierarchy and an example of a model with a methodological hierarchy. (Alternatively, describe one model with both, carefully distinguishing each component.)

   For the rest of the question, we focus specifically on substantive hierarchies in data. Regardless of modeling approach, why are hierarchical models useful in the field of Political Science? Name two examples of real or hypothetical data where a hierarchical approach would be beneficial and why.

   When modeling data with a hierarchical structure with *random effects*, we may use traditional methods or Bayesian methods. Name three advantages the Bayesian approach has over a traditional method.[1] Another possible strategy, particularly within traditional approaches, is to estimate a *fixed effects* model rather than a *random effects* model. Why might a researcher prefer one approach over the other, and how might the researcher's choice depend on other features of the data or model?

[1] Traditional methods include maximum likelihood and feasible generalized least squares. As a contrast to Bayes, you may discuss whichever type of technique you feel best about discussing, or both if you like.

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 kind of information you would hope to learn in your example and why it would be useful.

Afternoon Session: Analyzing Data

Answer one of the following two questions:

3. *Linear Regression:* Please analyze the survey data set *engagement.dta* using a linear regression model. The data set contains information on overall civic engagement during the 2012 presidential election. The variables are as follows (you must use them all):

   engscale Civic engagement scale (**dependent variable**; 0 = doesn't participate in any activity, 5 = participates in all activities).

   education Educational attainment (1 = didn't complete high school, 6 = post-graduate degree).

   income Household income (1 = less than $10,000, 18 = $250,000 or more).

   age Respondent's age.

   stghpid A measure of strength of partisanship (0 = Independent, 3 = strong party ID).

   ideodist Ideological distance to President Obama (0 = same position as President Obama, 6 = at the opposite extreme relative to President Obama).

   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?

   Please test the conditional hypothesis that educational attainment fosters civic engagement, and that the positive influence of education is especially intense for individuals with strong partisan attachments. 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.

   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 and influential data points—you may refer to the data points by row number as you do not know the names of survey respondents. Include the appropriate graphs or 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. *Logistic Regression:* Please analyze the survey data set *convenience.dta* using a logistic regression model.

   In recent years, states across the country have been introducing election reforms that make voting more convenient, including allowing citizens to vote by mail or in person before Election Day. The data set contains information on use of early voting opportunities in the 2012 general election.

   The outcome of interest is the variable **earlyvote**—whether the respondent voted early (either in person or by mail) or in person on Election Day. The predictors are:

   evlaws Dummy—whether the respondent lives in a sate that allows no-excuse absentee voting or no-excuse in person early voting.

   allbymail Dummy—whether the respondent lives in a state where everyone votes by mail ahead of Election Day (i.e. a state holding all-mail elections).

   education Educational attainment (1 = didn't complete high school, 6 = post-graduate degree).

   homeown Dummy—whether the respondent is a homeowner.

   campaign Dummy—whether the respondent was contacted by a political campaign.

   interest Interest in public affairs (1 = doesn't follow political news, 4 = follows political news most of the time).

   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 **evlaws** varies as a function of interest in public affairs. 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?