Methodology Minor Field Exam: Study Guide

Fall 2014

The methodology minor field exam will consist of two sessions. The morning question, which you will have three hours to complete, will require you to answer an essay question about the theory behind particular methods or the decisions facing analysts. The afternoon question, which you will have four hours to complete, will focus on applied data analysis, requiring you to fit a prescribed model using a provided data set and then compute additional quantities of interest (such as forecasts or model diagnostics).

The exam will be held in the Candler Hall computer lab, Room B04. This laboratory is equipped with all software you need to complete the exam: Microsoft Word, LaTeX (TeXshop on the Mac machines and TeXworks on the PC machines), R, and Stata. You may write your answer in either Word or LaTeX, as you prefer, and you may use any kind of data analysis software on the computer to complete the data analysis for the afternoon section. As a quick note: additional libraries for R or Stata are not necessarily installed, so be prepared to install any library or package you need as part of the process.

Also, during the afternoon section, when you complete your data analysis, you are free to use the Internet to look up anything you like as long as you do **not** attempt to contact any other person for help. In other words, if you use the Internet to download data, install packages, or look up documentation on commands, all of that is fine. You may not, however, post a question on a discussion board, send an e-mail to anyone, or otherwise contact anyone in any way. Your work must be your own. We recommend you review the processes you will need to use in either R or Stata ahead of time so that you can work more efficiently during the exam time, but consulting any static online reference is fine.

Morning Session: Statistical Theory and Modeling Decisions

Two of the following three questions will be on the morning portion of the exam, and you may choose which one of the two you wish to answer.

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:* Causal inference in the presence of a reciprocal relationship can be difficult. 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. Suppose, then, that you worked for a think tank that wanted to make a projection of how much per capita GDP would rise if a country expanded its democratic institutions enough to rise one point on the POLITY scale. It could be difficult to isolate the one-way effect of democracy on economic development if you believe there is a feedback loop between the two.

   Why it is so problematic to isolate a causal effect whenever there is a reciprocal relationship between variables? How prevalent of a problem is this in Political Science?

   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 four 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? What is an example of a real problem to which you would consider applying this technique?

   Consider the opening example of the reciprocal relationship between economic development and democratic institutions. As a policy analyst, how would you go about making your forecast to report back to policymakers? Be sure to describe both the kind of data and the estimation method you would use. Why will you stake your reputation on this research design? Do you have any reason for doubt in your forecast? Why or why not?

3. *Spatial Data Analysis:* Maps of data often show similarities between neighboring units. For instance, neighboring counties in the United States tend to have similar rates of lung cancer, and neighboring countries in the Arab world have engaged in similar numbers of protests since December 2010. We can see this simply by visual examination of the maps, or though computing test statistics on the raw variables of interest, such as Moran's $\mathcal{I}$ and Geary's $\mathcal{C}$.

What are **three** reasons we might see similarity among neighboring areal units on different variables of interest? For each of these theoretical causes of similarity, please answer the following two subquestions: What is the nature of the process that drives neighboring units to similar levels of the variable? When modeling a variable like this with spatially-referenced units, what are some of the available methods?

Turning back to the opening two examples: Which of these reasons is mostly likely to explain similarities among neighboring counties in lung cancer data? If you were modeling county-level lung cancer data, which method would you use? Why is this the best method?

What about protests during the Arab Spring? Which reason is most likely to explain similarities among these nations? If you were modeling the number of protests in these countries, which method would you use? Why is this the best method?

Afternoon Session: Analyzing Data

Two of the following three questions will be on the afternoon portion of the exam, and you may choose which one of the two you wish to answer. Each question that is asked will be accompanied by an appropriate data set located at a URL online. The question wording will be made more specific in the context of the real data, but the question form and skills required will remain the same.

4. Please analyze a data set using a logistic regression model. The outcome of interest is variable y, and the input variables are $x_1, x_2 \ldots, x_k$. 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?

   The model will call for an interaction term of some sort. 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?

   We will also propose a simpler specification of the model, and ask you to 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?

5. Please analyze a data set using a count model. The outcome of interest is variable y, and the input variables are $x_1, x_2 \ldots, x_k$. Your answer may lead you to report multiple versions of this model, so feel free to present the results from all of your models in

separate tables or in one big table, as you prefer. Start by fitting a Poisson model and reporting these results. Please test for overdispersion and zero inflation in these data. 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?

The model will call for an interaction term of some sort. 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.

6. Please analyze a time series data set using transfer function analysis. This data set includes two variables, y and x, though it may not be clear up front which variable is endogenous to the other.

   Begin by diagnosing the ARIMA process for y and for x, reporting your diagnosis and noise model estimates for each, and saving the prewhitened residuals for each. Show the initial plots of the autocorrelation (ACF) and partial autocorrelation (PACF) functions, and explain how you chose the ARIMA models you estimated. Demonstrate that the resulting prewhitened series are white noise using either a test statistic or a visual method.

   Please plot the cross-correlation function (CCF) of these two prewhitened series. Based on this plot, please explain which series causes the other and the proper specification of a transfer function (onset lag, numerator order, and denominator order).

   Next, report the results of a full model that includes the transfer function and the ARIMA noise model for the variable you identify as endogenous. Report the results of this model in a table including the coefficients, the standard errors, at least one fit statistic, and any additional information you would like. Please diagnose that that the residuals of this model are white noise and that there is no cross-correlation between the prewhitened exogenous variable and the transfer function residuals. (In the event that either test fails, please either respecify the model, or explain how you would if you start running short on time.)

   Finally, please plot the functional form of your estimated transfer function to show how the effect of a one-unit increase in the exogenous variable unfolds over time in the

endogenous variable. What can you determine from this substantive figure, as well as the statistical tests of the terms in the transfer function?