Political Methodology Comprehensive Examination, January 2015
Department of Political Science, George Washington University

Instructions: Read all questions before answering any of them. Answer 6 questions in total, with question 5, which will take longer, counting as two questions. Put differently, if you answer question 5, you only need to answer 5 questions in total. After the exam turn in an empirical paper demonstrating your ability to use statistical models OR schedule an oral exam.

1. Suppose a researcher is interested in how race affects turnout and runs a regression on individual level data with two independent variables, X_1 and X_2. X_1 is a dummy variable for African-Americans and X_2 is a dummy variable for Latinos. "Other race" is the reference category, and turnout is a dichotomous variable. The results of the regression return a coefficient of 2.36 on X_1, with a p-value = 0.01, and a coefficient of 1.89 on X_2, with a p-value = 0.07. Since X_1 is significant and X_2 is not, the researcher proclaims that we can be very confident that African-Americans have higher turnout than Caucasians, but cannot make such a statement for Latinos. Moreover, we can be very confident that African-Americans have higher turnout than Latinos. Are these conclusions valid? Can you suggest any other tests to validate these claims?

2. The January 2015 issue of *PS* contains an article entitled: "Can Big Data Solve the Fundamental Problem of Causal Inference?" How would you answer this question? In your answer, provide a definition of big data, since there is not a universally agreed upon definition of the concept.

3. You have estimated a cross sectional linear regression model when you realize that you have one percent battery life on your laptop and no way to recharge it, having left your power cord in your apartment. You have time to generate three regression diagnostic graphs and print them out to study. What graphs would you print, and how would you use them to assess the adequacy of your model?

4. Typical MLE/Nonlinear models classes cover several types of models where linear models can also be used. If your data has a dichotomous, ordered, or count dependent variable, why not just estimate both linear and nonlinear models, and if they yield similar results, present the linear models that more of your audience is likely to understand? What are the risks and benefits of this strategy, and under what circumstances is it a particularly bad idea?

5. Consider a bandit who is deciding whether to attack a village in two successive rounds. The bandit comes in two types: Strong and Weak. Weak types will lose for certain if they attack, whereas Strong types will win with probability p. Controlling the village in each round is worth W to the bandit and to the village (imagined as a unitary actor). An attack costs c for both.

The order of the game is as follows:

1. The bandit decides to attack the village or not. Payoffs are realized.

2. A bit later, the village decides whether to give up to the bandit before the next attack. If so, the bandit gains W, the village loses W without fighting, and the game ends.

3. If the village does not give up, the bandit again decides to attack the village or not.

Note that steps 2 and 3 happen regardless of the outcome in step 1. There is no discount factor between attacking rounds. Under what conditions is there a separating equilibrium in which the Strong type attacks in each round and the Weak type does not? What does the village do in step 2? (Hint: Be careful to check what is in the interest of each type given the village's action.)

6. Identify what you think are some of the key deficiencies in how political scientists have modeled "clustered data," both time-series cross-sectional data and multilevel data more generally. Identify what you believe are avenues and methods for enhancing the analysis of this type of data in order to maximize the types of substantive claims we are able to make.

7. Regression discontinuity(RD) research designs are increasingly popular in political science. What accounts for the consensus that RD designs provide more credible claims to uncovering causal effects than matching designs? If you used a RD design, what diagnostics would you undertake to evaluate whether the discontinuity was plausibly random? Further, what threats to external validity may RD designs introduce and how should they be addressed by researchers?

8. In *Quasi-Experimentation,* Cook and Campbell wrote that "randomization is the great ceteris paribus" that enables us to claim that the groups we compare are equivalent, giving us strong claims to internal validity. Setting aside questions of generalizability (external validity), however, what potential problems can arise in the implementation of field experiments that can threaten internal validity? How can these problems be addressed?

9. Network analysis includes, among other things, a set of measurement tools. Give examples of two of these tools. Explain the theoretical basis for these tools, what advantages they have over more conventional measurement tools, and their limitations. For each tool, give an example of how one might use it in applied research, including an explanation of why it would be useful in this context and how it can improve our inferences.