# Methodology Minor Field Exam

Fall 2020

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.

For the morning session, internet usage is prohibited. For the afternoon session, you may use the internet to download software packages or look up reference information as you complete the data analysis. Your work must be fully your own. Enjoy this opportunity to showcase your skills.

## Morning Session: Statistical Theory and Modeling Decisions

Answer one of the following two questions:

1. *Causal Inference:* The tools of causal inference are often used to determine how effective a certain policy is. For example, a causal study might evaluate how a universal health insurance program in Mexico affected citizens' reporting access to a primary care doctor. Presumably, for the policy to be deemed a success, the policy would have to increase the number of citizens reporting access to a primary care doctor.

   Why, theoretically, is it difficult to determine how effective a policy is in shaping a public outcome of interest?

   Name three methods that are used for causal inference. Describe how each method is applied in practice and why it theoretically helps us isolate a treatment effect. For each method give an example of a hypothetical policy, a hypothetical outcome of interest, and how you could use that method to estimate the effect that the policy had on the outcome of interest. (If it is easier to use real policies and outcomes in your examples, feel free.) For the sake of each example, you may dictate how the policy is rolled-out, whether there were temporal observations before and after the policy roll-out, whether there is geographic variation, and whether there is individual-level variation. Possible examples might be policies regarding employment, election administration, policing, auditing, trade practices, or anything else that may come to your mind.

2. *Linear Model Theory:* When estimating a linear model via OLS, one assumes the Gauss-Markov assumptions to be true in order for the estimates to be BLUE. What are these Gauss-Markov assumptions? What does it mean for an estimator to be BLUE? What are the most common violations of Gauss-Markov assumptions and what types of data are most likely to lead to violations of these assumptions? What are the implications to violating Gauss-Markov, specifically in terms of interpreting model results? Finally, which assumption do you feel is the most important to not violate and why?

# Afternoon Session: Analyzing Data

Answer one of the following two questions:

3. *Poisson Regression:* Please analyze the dataset *couart2.dta* using a poisson regression model. The dataset contains information on the number of articles published by PhD students during the last 3 years of their education. The variable of interest, **art**, is the number of articles published by students in the last 3 years they attend a PhD program.

   The input variables (you must use them all) are:

   **fem** : Dummy for gender (1 = female).

   **mar** : Dummy for marital status (1=married).

   **kid5** : Number of children

   **phd** : Prestige score of PhD granting institution (higher = more prestigious).

   **ment** : Number of articles published by student's mentor in last 3 years.

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

   Present graphs of predicted counts against all covariates.

   Finally, discuss whether or not you think Poisson regression is the appropriate technique for these data and justify your answer. If no, then discuss other options and why they may be more appropriate.

4. *Linear Regression:* Please analyze the dataset *Nicaragua2017_comps.dta* using a linear regression model.

   The dataset contains information about vote buying and trust in elections in Nicaragua shortly before the 2017 legislative elections. The variable of interest, **b47a**, is a Likert scale measuring respondents' expressed trust in Nicaraguan elections (dependent variable; 1=does not trust elections at all, 7=strong trust in elections).

   The input variables are as follows (you must use them all):

   **rural** : Rural (1=rural locale, 0=urban locale).

   **mujer** : Woman (1=female, 0=male).

   **edad** : Age (1=18-25, 6=65+).

   **edr** : Level of Education (1=no education, 4=post-secondary education).

   **quintall** : Wealth Quintiles (1=poorest, 5=wealthiest).

   **direct exposed** : Offered Benefit for Vote (1=was asked to sell vote, 0=was not asked to sell vote).

   **approving norm** : Approving Norm (scale; 0=strong disapproval of vote buying exchanges, 1=strong approval of vote buying exchanges).

   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 being asked to sell one's vote has a negative effect on trust in elections, and that this effect is especially strong among those who disapprove of vote buying. 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. 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? Assess whether or not there are problems with collinearity and heteroskedasticity. 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?