Quantitative Empirical Methods Exam

Yale Department of Political Science, August 2018

You have seven hours to complete the exam. This exam consists of three parts.

Back up your assertions with mathematics where appropriate and show your work. Good answers will provide a direct answer that illustrates an understanding of the question, and calculations or statistical arguments to validate the answer. Where applicable, exceptional answers will include all of these *as well as* proofs that are technically complete, including formally articulating sufficient assumptions and regularity conditions. Questions will not be weighted equally. A holistic score will be assigned to the exam. Therefore, it is important to demonstrate your understanding of the material to the best of your ability.

Part 1 (Short Answer Section) consists of five short answer questions. *Advice*: Note there are multiple correct answers to some questions. We encourage you to give the most complete (but still succinct) solution possible. Do not leave sub-parts of questions unanswered.

Part 2 (Essay Section) contains a recent, well-regarded empirical article. We will ask you to offer an evaluation of its methodological approach and presentation of results. In particular, we will advise you to pay particular attention to the identification conditions (either explicit or implicit), the associated estimation strategy, and possible threats to inference. Your response may be anywhere from 500 to 1500 words.

The only aids permitted for Parts 1 and 2 are (i) one page of double-sided notes, (ii) a word processor on one of the Statlab computers to write up your answers (you may also write up your answers to Part 1 using pencil/pen and paper). After handing in your answers for Parts 1 and 2 of the exam, you may begin Part 3 (though feel free to look ahead). You may hand in Parts 1 and 2 whenever you wish, but we recommend spending no longer than five hours on Parts 1 and 2.

Part 3 (Computer Assisted Section) will involve using statistical software to answer one longer exercise with five associated questions. A complete answer to Part 3 will include code and output, as well as your written answers. *Advice*: We recommend that you explain what you are trying to do in comments in your code. Even if you are not able to execute your program correctly, you can receive partial credit for explaining clearly what you wanted to do and why.

For Part 3, you are permitted access to any and all written materials, as well as (i) unrestricted use of your own computer with access to the internet or (ii) use of a Statlab computer with access to the internet. The only restriction for Part 3 is that you may not interact with any person, online or otherwise. For Part 3 (Computer Assisted Portion) of the exam, please turn in a hard copy of your code to Colleen, and also email a digital copy of the code to colleen.amaro@yale.edu.

1 Short Answer Section

1. Is statistical significance "transitive" in the sense that if A is statistically significantly different from B and B is statistically significantly different from C, A is statistically significantly different from C? If Yes, explain why. If No, give a counterexample.

2. There is a variable, Y, which changes value over time. You estimate three models using this data.

   Model 1 is $Y = A + BT + \epsilon$, a standard Ordinary Least Squares model.

   Model 2 is $Y = \alpha + \beta_1 T + \beta_2 D + \epsilon$ where D is a binary variable that takes on the value 1 if time is greater than or equal to some cutoff c, and takes on the value 0 if time is less than said cutoff c.

   Model 3 is a regression discontinuity design, with a discontinuity at c.

   (a) Suppose β_2 is estimated to be large and positive. What will be the consequences for Model 1? Consider the following questions, and be as precise as possible. What do you know about how well Model 1 fits the data? What do you know about the relationship between α and A? What do you know about the relationship between β_1 and B?

   (b) What are the differences between Model 2 and Model 3? Be specific, focusing especially on assumptions and interpretation.

3. Assume that the conditional expectation function of Y given X and Z is $E[Y|X, Z] = 1+3XZ+X^2$. Further assume that X and Z are independent and each distributed according to the standard uniform distribution U(0, 1).

   (a) What is the marginal effect of X on Y when X = 0 and Z = 1?

   (b) What is the marginal effect of Z on Y when X = 0 and Z = 1?

   (c) What is the marginal effect of X on Y when both X and Z are at their means?

   (d) What is the average marginal effect of X on Y ?

4. Consider an experiment with a finite population of 6 people who live on the same street, indexed $i = 1, 2, 3, 4, 5, 6$. Exactly 3 people are assigned to treatment by complete random assignment, with the remainder in control. Assume the treatment assignment is the only source of randomness.

   (a) Do not assume noninterference. How many potential outcomes can each subject express?

   (b) Assume noninterference. How many potential outcomes can each subject express?

   (c) Assume that subject i can only interfere with subject j if $|i - j| = 1$ (i.e., they are neighbors). How many potential outcomes can each subject express?

5. Compare the assumptions underlying the logit model to those underlying the linear probability model. What are the most important differences between the two? Under what conditions should a researcher choose to run a logit? Why? What are the downsides of a linear probability model? Are there any circumstances under which a researcher should choose to run a linear probability model?

2 Essay section

Please read the attached article: Malesky, Edmund, Paul Schuler and Anh Tran. 2012. “The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in An Authoritarian Assembly.” American Political Science Review 106(04):762 - 786.

Evaluate and critique the paper. Offer a critical evaluation of its methodological approach and presentation of results. Note: “critical” does not imply that you must only criticize—you should give credit to the authors when their arguments are convincing and/or novel with respect to standard practice. Your response may be anywhere from 500 to 1500 words.

We advise you to pay particular attention to the identification conditions (either explicit or implicit), the associated estimation strategy, and possible threats to inference. Justify each of your claims and, where applicable, suggest ways in which this line of research might be improved. (We do not expect you to have special expertise in the topic area, but we do expect you to bring to bear your general analytical skills as a political scientist).

Please focus closely on the following questions.

1. What is the effect of transparency on performance? How is this estimated? What assumptions underlie this estimate? Do these assumptions seem valid here?

2. How does “internet penetration” change the effect of transparency on performance? How is this estimated? What assumptions underlie the estimation of this interaction effect? Do these assumptions seem valid here?

3. What is the effect of transparency on performance when internet penetration is very high? How is this estimated? What assumptions underlie the estimation of this effect? Do these assumptions seem valid here? (Hint: use the information provided in the paper to estimate the effect of transparency on performance when internet penetration takes on a value of 6. How confident are you in this estimate?)

3 Computer Assisted Portion

In collaboration with a government agency from a developing country, a researcher conducted a pilot study to evaluate if higher wages attract a larger and more qualified pooled of applicants. Two different wage offers were randomly assigned across 100 sites. In one half of the sites, a salary of $250 USD per month was offered. In the other half, a salary of $150 USD per month was offered. The government agency collected information to measure an array of candidate characteristics, including years of schooling.

We would like you to analyze the data from this study following the instructions below. Please answer all questions. You may use the statistical software of your choice. Turn in your code and output. Please insert comments on your code explaining what you intent to do, and responding to the questions. (We recommend that you explain what you are trying to do in comments. Even if you are not able to execute your program correctly, you can receive partial credit for explaining clearly what you wanted to do and why.)

You can download the data here https://www.dropbox.com/s/7mvwqndhcye5s38/data.csv?dl=0.

The data set contains four variables. ID is a unique identifier per site; Z takes the value of 1 for sites randomly assigned to a high wage offer, and 0 otherwise; D takes the value of 1 for sites where a high wage was offered, and 0 otherwise; Y is the average years of schooling of candidates who applied for the position per site.

1. Inspect the quality of the data. Verify the data is clean and ready to use, then compute and report the descriptive statistics you consider relevant to the study.

2. Assume for now that the study had perfect compliance with treatment assignment (i.e. all units that were assigned to be treated with the high wage offer were treated with the high wage offer, and all units assigned to be treated with the low wage offer were treated with the low wage offer). Since the sample size of this study is 100 sites, you will start your analysis asking very little from the data with a non-parametric hypothesis test. For the outcome variable Y test the sharp null hypothesis of no effect on any unit against the alternative of some effect using a signed-rank test. Compute p-values using randomization inference. Discuss your results.

3. Calculate the $\widehat{ITT_D}$ (i.e. the effect of assignment to treatment on treatment), $\widehat{ITT}$ (i.e. the reduced form effect), and $\widehat{CACE}$ (i.e. the Complier Average Causal Effect). Compute p-values using randomization inference. Discuss your results.

4. You have one instrument (i.e. random assignment to treatment), and one endogenous regressor (i.e. treatment delivered). Is it right to say that your just-identified IV estimate is unbiased? Would your answer be different if the sample size of the study was 1000? Discuss your answer.

5. The government agency assigned sites to the two experimental conditions using a public lottery. The researcher is worried that the lottery itself could impact outcomes directly, which poses a challenge to the exclusion restriction. Devise a sensitivity test that helps you assess the robustness of your results to deviations from the assumption that the exclusion restriction holds exactly.