Quantitative Empirical Methods Exam

Yale Department of Political Science, August 2019

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 seven 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 (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 anyone, 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. Suppose there is a population of 500 subjects, where 300 of the subjects are Democrats and 200 of the subjects are Republicans. You randomly sample one person from this population, and record whether or not this person was a Democrat or a Republican. Formally represent this random generative process as a probability space.

2. Provide brief definitions of the following terms:

   (a) standard deviation

   (b) standard error

   (c) statistical significance

   (d) confidence interval

   (e) p-value

3. Assume that the conditional expectation function of Y given X and Z,

   $$\mathrm{E}\left[Y|X,Z\right] = 10 + 8XZ + 7X^2.$$

   Further assume that X and Z are independent and each distributed according to the standard uniform distribution $U(0,1)$.

   (a) What is $\mathrm{E}\left[Y|X=1,Z=0\right]$?

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

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

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

4. Assume that you have n observations i.i.d. $X \sim N(\mu,1)$. Prove that the ML estimate of μ is the sample mean, $\overline{X} = \sum_{i=1}^{n} X_i$. Recall that the normal PDF,

   $$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \forall x \in \mathbb{R}.$$

5. Consider a sharp regression discontinuity design with outcome Y, treatment D and forcing variable X (with a cutpoint at zero). Suppose that you have a consistent estimator of

   $$\theta = \lim_{x \to 0^+} \mathrm{E}\left[Y|D=1,X=x\right] - \lim_{x \to 0^-} \mathrm{E}\left[Y|D=0,X=x\right].$$

   Denote this estimator $\hat{\theta}$.

   (a) Articulate a set of (nontrivial) conditions under which $\hat{\theta}$ is consistent for a causal effect (of D on Y).

   (b) In what ways do scholars seek to validate the presence of these conditions in empirical research? How persuasive are these efforts?

6. Suppose that D_i and Z_i are binary. Let

    - $\mathrm{E}[Y_i|D_i = 1, Z_i = 1] = 1.00$,
    - $\Pr[D_i = 1, Z_i = 1] = 0.25$,
    - $\mathrm{E}[Y_i|D_i = 0, Z_i = 1] = 0.50$,
    - $\Pr[D_i = 0, Z_i = 1] = 0.50$,
    - $\mathrm{E}[Y_i|D_i = 0, Z_i = 0] = 0.50$,
    - $\Pr[D_i = 0, Z_i = 0] = 0.25$,
    - $\Pr[D_i = 1, Z_i = 0] = 0.00$.

    Also consider the following assumptions:

    (i) $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$.

    (ii) $D_i = D_i(1)Z_i + D_i(0)(1 - Z_i)$.

    (iii) $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i$.

    (iv) $(Y_i(0), Y_i(1), D_i(1), D_i(0)) \perp\!\!\!\perp Z_i$.

    (v) $\mathrm{Supp}[Y_i(d)] \subseteq [0, 1]$, for $d \in \{0, 1\}$.

    You do *not* need to simplify fractions or use any computation – e.g., $\frac{0.50}{1.00+0.25}$ is a perfectly acceptable answer.

    (a) Under assumptions (i) and (v), compute sharp bounds for $\mathrm{E}[Y_i(1) - Y_i(0)]$.

    (b) Under assumptions (i) and (iii), compute $\mathrm{E}[Y_i(1) - Y_i(0)]$.

    (c) Under assumptions (i), (ii) and (iv), compute $\Pr[D_i(1) > D_i(0)]$, $\Pr[D_i(1) = D_i(0) = 1]$, $\Pr[D_i(1) = D_i(0) = 0]$, and $\Pr[D_i(1) < D_i(0)]$.

    (d) Under assumptions (i), (ii) and (iv), compute $\mathrm{E}[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)]$.

7. For the analysis of longitudinal data (i.e., TSCS or panel data), there is debate in the social sciences about the use of fixed effects, random effects, and pooled regression for estimating causal effects. What are fixed effects regression, random effects regression, and pooled regression? Summarize and critically assess the arguments made about these types of estimators.

2 Essay section

Read the article attached to your exam. Offer a critical evaluation of its methodological approach and presentation of results. Note: “critical” does not imply that you must only criticize – it is recommended that you give credit to the authors if and 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.)

Article: Hall, Andrew B., Huff, Connor, & Kuriwaki, Shiro (2019). Wealth, Slaveownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War. *American Political Science Review*, 113(3), 658–673.

3 Computer Assisted Portion

Suppose the researcher, Professor Robo, collects n measurements on one outcome Y_i, a randomized treatment D_i, and a set of pretreatment covariates $\mathbf{X}_i = \{X_{(1)i}, X_{(2)i}, ..., X_{(K)i}\}$, all of which are binary. Robo correctly assumes SUTVA, with $Y_i(0)$ and $Y_i(1)$ denoting potential outcomes. Loosely speaking, Robo will use statistical methods to determine whether or not there are subgroups such that the average effect of treatment is nonzero.

You are going to conduct a simulation to assess the operating characteristics of Robo's statistical methodology, by simulating both the data and how Robo would analyze it. As the designers of this simulation, we will assume that all observable variables are generated by nature as independent Bernoulli random variables with $p = .5$ (although this is unknown to Robo).

Additional instructions: For all of your Monte Carlo calculations, use at least 1000 simulations. For all confidence intervals and significance tests, use $\alpha = 0.05$. Figures must not have haphazard axes.

1. What is the true value of $\mathrm{E}\left[Y_i(1) - Y_i(0)\right]$ in this setting? Are there any values of $\mathbf{x}_i$ such that the true $\mathrm{E}\left[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}_i\right] \neq 0$? (This question does not require any simulation.)

2. Robo's first step is to test whether there are any main effects. Robo uses the difference-in-means estimator, $\hat{\mathrm{E}}\left[Y_i|D_i = 1\right] - \hat{\mathrm{E}}\left[Y_i|D_i = 0\right]$, to estimate $\mathrm{E}\left[Y_i(1) - Y_i(0)\right]$. Robo weds this with the associated SE estimator,

$$\sqrt{\frac{\widehat{\mathrm{Var}}\left[Y_i|D_i = 1\right]}{\sum_{i=1}^n D_i} + \frac{\widehat{\mathrm{Var}}\left[Y_i|D_i = 0\right]}{\sum_{i=1}^n (1 - D_i)}}$$

to calculate normal approximation-based confidence intervals and two-tailed p-values.

   (a) For each $n \in \{50, 100, 500, 1000\}$, present a histogram of the sampling distribution of the difference-in-means estimator of $\mathrm{E}\left[Y_i(1) - Y_i(0)\right]$.

   (b) For each $n \in \{50, 100, 500, 1000\}$, calculate the probability that Robo rejects the null hypothesis that $\mathrm{E}\left[Y_i(1) - Y_i(0)\right] = 0$.

   (c) Comment on the properties of the difference-in-means estimator and associated confidence intervals for main effects in this setting.

3. Robo decides to look at the subgroup effect for a single, prespecified subgroup. Define the subgroup effect for subjects such that covariate j is equal to k as

$$\theta_{(j)k} = \mathrm{E}\left[Y_i(1) - Y_(0)|X_{(j)i} = k\right], \forall j \in \{1, ..., 10\}, k \in \{0, 1\}.$$

Robo considers plug-in estimators of the form:

$$\hat{\theta}_{(j)k} = \hat{\mathrm{E}}\left[Y_i|D_i = 1, X_{(j)i} = k\right] - \hat{\mathrm{E}}\left[Y_i|D_i = 0, X_{(j)i} = k\right], \forall j \in \{1, ..., 10\}, k \in \{0, 1\},$$

each with an associated SE estimator,

$$\sqrt{\frac{\widehat{\mathrm{Var}}\left[Y_i|D_i = 1, X_{(j)i} = k\right]}{\sum_{i=1}^n D_i \mathrm{I}\left[X_{(j)i} = k\right]} + \frac{\widehat{\mathrm{Var}}\left[Y_i|D_i = 0, X_{(j)i} = k\right]}{\sum_{i=1}^n (1 - D_i) \mathrm{I}\left[X_{(j)i} = k\right]}},$$

that Robo uses to form normal approximation-based confidence intervals and two-tailed p-values. Regardless of the outcome of the experiment, Robo restricts attention to a single subgroup: subjects such that covariate 1 is equal to 0.

(a) In words, how do you compute $\hat{\theta}_{(1)0}$?

(b) For each $n \in \{50, 100, 500, 1000\}$, present a histogram of the sampling distribution of $\hat{\theta}_{(1)0}$.

(c) For each $n \in \{50, 100, 500, 1000\}$, calculate the probability that Robo rejects the null hypothesis that $\theta_{(1)0} = 0$.

(d) Comment on the properties of $\hat{\theta}_{(1)0}$ and associated confidence intervals for this subgroup effect in this setting.

4. Robo now decides to engage in exploratory research to search for a significant subgroup effect. Robo uses the following procedure:

- For all 20 subgroups indexed $j \in \{1, ..., 10\}, k \in \{0, 1\}$, Robo computes $\hat{\theta}_{(j)k}$ and an associated normal approximation-based p-value, denoted $p_{(j)k}$.
- Robo considers the subgroup with the lowest associated p-value. (If not unique, Robo will select randomly among minimizers.) Let $J, K = \arg\min_{j,k} p_{(j)k}$, noting that J, K are random variables and change according to the random draw of the data.
- Robo then reports $\hat{\theta}_{(J),K}$, or the difference-in-means estimator as applied to the group with the smallest p-value, and $p_{(J)K}$, the smallest p-value.

(a) For each $n \in \{50, 100, 500, 1000\}$, present a histogram of the sampling distribution of $\hat{\theta}_{(J),K}$.

(b) For each $n \in \{50, 100, 500, 1000\}$, calculate the probability that $p_{(J)K} < 0.05$.

(c) Comment on the implications of your simulation results for applied research.