Quantitative Empirical Methods Exam

Yale Department of Political Science, August 2017

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, and a holistic score will be assigned to the exam, and thus 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, and we encourage you to give the most complete (but still succinct) solution possible. Do not leave sub-parts of questions unanswered. Starred subquestions are especially difficult.

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 three 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. 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

Starred subquestions are especially difficult, and may be skipped. (*) indicates difficult. (**) indicates very difficult.

1. Prove or disprove the following claims.

   (a) If $\mathrm{Var}[XY] > 0$, then both $\mathrm{Var}[X] > 0$ and $\mathrm{Var}[Y] > 0$.

   (b) If $\mathrm{Var}[XY] > 0$, then both $\mathrm{E}[X^2] > 0$ and $\mathrm{E}[Y^2] > 0$.

   (c) If $\mathrm{Var}[XY] > 0$, then either $\mathrm{E}[X] \neq 0$ or $\mathrm{E}[Y] \neq 0$.

   (d) If both $\mathrm{Var}[X] > 0$ and $\mathrm{Var}[Y] > 0$, then $\mathrm{Var}[XY] > 0$.

2. Suppose that you collected 100 observations that are i.i.d. X. Of these, 80 observations take on the value 0, and 20 observations take on the value 1. Estimate a 95% confidence interval for $\mathrm{E}[X]$ under a normal approximation.

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

$$\mathrm{E}[Y|X, Z] = 1 + 3XZ + X^2.$$

   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. Assume you have $n \times k$ regressor matrix $\mathbf{X}$ and n-length outcome vector $\mathbf{Y}$. (Treat $\mathbf{Y}$ as a $n \times 1$ matrix.) Assume $\mathbf{X}$ contains a constant.

   (a) Under what conditions is the OLS solution from the regression of $\mathbf{Y}$ on $\mathbf{X}$ uniquely defined? (I.e., there exists one and only one possible solution that minimizes the sum of squared residuals.) Consider an OLS solution $\hat{\boldsymbol{\beta}}$.

   (b) Give an example of data for which the OLS solution would not be uniquely defined.

   (c) Assuming that there exists a uniquely defined OLS solution, write down a closed-form expression for $\hat{\boldsymbol{\beta}}$ using matrix algebra.

   (d) Derive your result in (c). (**)

   (e) Assuming that there exists a uniquely defined OLS solution, what is the mean of the elements of $\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$?

   (f) Prove your answer in (e). (*)

(g) Suppose that the OLS solution is not uniquely defined. Does your result for (e) hold for *any* $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared residuals? (**)

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

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

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). What qualifications on this causal effect might you place? (I.e., is the subpopulation it applies to special in any way?)

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

6. Consider a randomized audit experiment in which bureaucrats are sent emails from putatively White or Latino citizens asking for voting information. The researchers are interested in both the effect of the Latino name on the email response rates of bureaucrats, and on the "tone" of their email responses. The treatment Z (putatively Latino names) is randomized by a simple coin flip (1 if White, 0 if Latino). The outcome R indicates whether or not the researchers received a response (1 if yes, 0 if not). The outcome Y indicates whether the tone of the reply email was "friendly" or not (1 if friendly, 0 if not), taking on the value of -99 if $R = 0$. Suppose that SUTVA holds for both R and Y with respect to Z.

(a) Prove that, if Z has no effect on R whatsoever (i.e., $\Pr[R(0) = R(1)] = 1$), then

$$\mathrm{E}\left[Y|Z=1, R=1\right] - \mathrm{E}\left[Y|Z=0, R=1\right] = \mathrm{E}\left[Y(1) - Y(0)|R=1\right].$$

(b) Prove that, if Z is allowed to have an effect on R, then it it is possible that

$$\mathrm{E}\left[Y|Z=1, R=1\right] - \mathrm{E}\left[Y|Z=0, R=1\right] \neq \mathrm{E}\left[Y(1) - Y(0)|R=1\right].$$

(c) Discuss the implications of your finding for practice.

7. 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? (*)

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 should only criticize — it is recommended that you 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).

Article: Jens Hainmueller, Dominik Hangartner, and Giuseppe Pietrantuono. 2017. Catalyst or Crown: Does Naturalization Promote the Long-Term Social Integration of Immigrants? *American Political Science Review*.

3 Computer Assisted Portion

Assume that $\log[f(y|\theta, x)] = c\left(|y - \theta x|^2 + |y - \theta|\right)$ for some constant $c < 0$.

1. Using the dataset below, graph the likelihood with respect to θ.
2. Numerically compute the MLE $\hat{\theta}$.
3. Estimate the standard error of the MLE using the bootstrap.
4. Bonus: Compute the model-based standard error. Does the disagreement between the bootstrap-based standard error and model-based standard error tell you anything?

x	y
2	8
-14	-13
-3	-11
16	11
-7	-23
20	38
9	12
21	28
19	8
-4	-4
10	23
-17	-6
2	-10
15	49
-1	-8
13	9
6	1
8	13
19	8
-5	-6