Quantitative Empirical Methods Exam

Yale Department of Political Science, August 2016

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 five 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.

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 six 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

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

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

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?

2. Suppose that the data generating process is i.i.d. $Y_i = a + bX_i + u_i$, with $\mathrm{E}\left[u_i|X_i\right] = 0$, $\Pr(X_i < 0) \in (0,1)$, and $\mathrm{Var}\left(X_i\right) > 0$. The researcher observes only values of X_i when $X_i \geq 0$. Suppose that the researcher drops all observations of (Y_i, X_i) with missing values on X_i, and performs an ordinary least squares (OLS) regression on the remaining values of Y_i on X_i (and a constant). Will the estimated slope from this regression generally be consistent for b? Why or why not?

3. Suppose that you have n mutually independent draws from a normally distributed random variable X with known variance $\mathrm{Var}\left(X\right) = 1$, and unknown mean $\mathrm{E}\left[X\right] = \mu$.

   (a) What is the maximum likelihood estimate of μ? Denote this $\hat{\mu}$.

   (b) What is $\lim_{n\to\infty} n\mathrm{Var}\left(\hat{\mu}\right)$?

   (c) What is $\lim_{n\to\infty} n\mathrm{E}\left[\hat{\mu} - \mu\right]$?

   (d) What is $\lim_{n\to\infty} n\mathrm{E}\left[(\hat{\mu} - \mu)^2\right]$?

   (e) Suppose that $\mathrm{Var}\left(X\right)$ were unknown. Would the maximum likelihood estimate of μ differ without knowledge of the variance, $\mathrm{Var}\left(X\right)$?

   (f) *Bonus.* Suppose that we know that μ is an integer. (You may assume $\mathrm{Var}\left(X\right) = 1$.) Propose an estimator that is asymptotically more efficient than $\hat{\mu}$. You may choose your own definition of asymptotic efficiency. [If you cannot give a formal proof, give an intuition.]

4. Suppose n observations are taken i.i.d. from (Y, D, Z, X), where Y, D, Z, and X are (scalar) random variables. Consider the following procedure:

   - The researcher estimates the following model using OLS: $D = \alpha_0 + \alpha_1 Z + \alpha_2 X + U_1$. Denote the estimated coefficients from this regression as $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$.
   - Denote $\hat{D} = \hat{\alpha}_0 + \hat{\alpha}_1 Z + \hat{\alpha}_2 X$.
   - The researcher then estimates the following model using OLS: $Y = \beta_0 + \beta_1 \hat{D} + \beta_2 X + U_2$. Denote the estimated coefficients from this regression as $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$.

Articulate a set of nontrivial conditions under which $\hat{\beta}_1$ is consistent for a causal effect of D on Y.

5. Suppose that we are trying to conduct inference on $\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right]$. Assume that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right]=\mu$, such that μ is finite. Let $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i$. Further suppose we computed a Wald-type confidence interval as $\hat{\mu}\pm\frac{1.96}{n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2}$. Give at least one example of a data generating process such that the asymptotic coverage of the resulting confidence interval will not be (at least) 95%.

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 – where praise is warranted, or where the authors' claims are well-justified, 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: Hainmueller, Jens and Dominik Hangartner. 2013. Who gets a Swiss passport? A natural experiment in immigrant discrimination. *American Political Science Review*. 159–187.

3 Computer Assisted Portion

Consider the following data generating process:

$$Y_i = \alpha + \beta X_i + u_i.$$

Assume that X_i is binary with $\Pr(X_i = 1) = 1/4$ and $\mathrm{E}\left[u_i | X_i\right] = 0$.

The researcher observes n i.i.d. draws from (Y_i, X_i). Suppose that the researcher fits an ordinary least squares (OLS) regression of Y_i on X_i and a constant. Denote the estimated coefficient on X_i as $\hat{\beta}$.

We will ask you to conduct a series of simulation studies to assess the behavior of Wald-type normal approximation-based confidence intervals for $\hat{\beta}$ constructed using "classical" standard errors, "robust" standard errors and the bootstrap. So as to guarantee non-collinearity in estimation, as you proceed, condition on the event that $0 < \sum_{i=1}^{n} X_i < n$. (Throughout, note that we are asking for Wald-type normal approximation-based confidence intervals, so do not use the percentile bootstrap.)

Use at least 1000 simulations and at least 500 bootstrap replicates in computing your answers.

1. Assume that $u_i = U(-1, 1)$, where $U(a, b)$ denotes the uniform distribution over the interval $[a, b]$.
   (a) Suppose that $n = 10$. What is the coverage of Wald-type 95% confidence intervals for $\hat{\beta}$ using (i) classical (OLS), (ii) robust (Huber-White), and (iii) bootstrap standard errors?
   (b) How about for $n = 100$? Repeat for (i)-(iii).
   (c) And $n = 2500$? Repeat for (i)-(iii).
2. Assume that $u_i = \begin{cases} U(-1, 1) & : X_i = 0 \\ U(-3, 3) & : X_i = 1 \end{cases}$, where $U(a, b)$ denotes the uniform distribution over the interval $[a, b]$.
   (a) Suppose that $n = 10$. What is the coverage of Wald-type 95% confidence intervals for $\hat{\beta}$ using (i) classical (OLS), (ii) robust (Huber-White), and (iii) bootstrap standard errors?
   (b) How about for $n = 100$? Repeat for (i)-(iii).
   (c) And $n = 2500$? Repeat for (i)-(iii).
3. Throughout,
   (a) did your answers depend on the true values of α or β? Why or why not?
   (b) did you find any differences between the results in Q1 and Q2? Why or why not?
   (c) did your answers depend on n? Were some procedures preferable to others with small n? What practical conclusions might you draw?