Quantitative Empirical Methods Exam

Yale Department of Political Science, August 2014

You have seven hours to complete the exam. This exam consists of two parts. Provide complete answers to **both** sections.

Back up your assertions with algebra wherever appropriate, and remember to 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.

Part 1 (Short Answer Section) consists of ten short answer questions. The only aids permitted for Part 1 are (i) one page of double-sided notes, (ii) a calculator, and (iii) a word processor on one of the Statlab computers to write up your answers (you may also write up answers using pencil/pen and paper). *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.

After handing in your answers for Part 1 of the exam, you may begin Part 2 (though feel free to look ahead). You may hand in Part 1 whenever you wish, but we recommend spending no longer than five hours on Part 1.

Part 2 (Computer Assisted Section) will involve using statistical software to answer one longer exercise with six associated questions. A complete answer to Part 2 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 2, 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 2 is that you may not interact with anyone, online or otherwise.

For Part 2 (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,

$$\mathrm{E}\left[Y|X,Z\right] = a + bX + cZ + dXZ + eZ^2 + fX^2.$$

What is the "marginal effect" of X on Y (i.e., instantaneous change in the conditional expectation function) when $X = x$ and $Z = z$?

2. Suppose that you collected 100 observations that are i.i.d. X. Exactly 50 observations take on the value 0. Exactly 50 observations take on the value 10. Estimate a 95% confidence interval for $\mathrm{E}\left[X\right]$ under a normal approximation.

3. (a) Loosely speaking, what is the law of large numbers?

   (b) Provide a formal statement of the weak law of large numbers.

   (c) Provide a proof of the weak law of large numbers. You may assume that all moments are finite and you may use well-known mathematical results (e.g., inequalities).

   (d) How does the weak law of large numbers differ from the strong law of large numbers?

4. (a) What is a confidence interval?

   (b) What is a posterior interval (a.k.a., Bayesian confidence interval, credible interval)? How does a posterior interval differ from a confidence interval?

   (c) What do these interval concepts imply about the differences between frequentist and Bayesian statistical inference?

5. Suppose that you have n observations i.i.d. (Y, D, Z). Articulate a set of (nontrivial) conditions under which the Wald IV estimate, $\widehat{\mathrm{Cov}}\left(Y,Z\right)/\widehat{\mathrm{Cov}}\left(D,Z\right)$, well approximates a causal effect (of D on Y). $\widehat{\mathrm{Cov}}\left(.\right)$ denotes the sample covariance.

6. Suppose that you observe n independent coin flips from a (potentially biased) coin that returns 1 with probability p and 0 with probability $1 - p$. Denote the vector of observed values of the coin flips as $(Y_1, ..., Y_n)$.

   (a) Provide a closed-form expression for the maximum likelihood estimate of p. Denote the estimator $\hat{p}$. (You do not need to derive the estimator; just provide the expression.)

   (b) Explain why the following four statements are true:

      i. $\hat{p}$ is unbiased.

      ii. $\hat{p}$ is consistent.

      iii. If $0 < p < 1$, then $\hat{p}$ is asymptotically normal.

      iv. If $0 < p < 1$, then $\hat{p}$ is root-n consistent.

(c) Suppose that you hold a strong belief that p is either at or very close to 0.5. (You may formalize "strong belief" however you wish.) In light of this belief, propose an estimator that might improve on $\hat{p}$. Justify this estimator.

(d) Professor Smedley claims, "Maximum likelihood estimates are efficient, and thus your estimator cannot improve on $\hat{p}$." Explain why Smedley is mistaken, and why your estimator might improve on $\hat{p}$.

7. Consider a random vector $(X_1, X_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (1, 1)$ and $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

   (a) Sketch the marginal PDF and marginal CDF of X_1.

   (b) Give an approximate value for $\Pr(X_1 < 2.96 | X_2 = \sqrt{\pi - 1.2})$.

8. A political scientist conducts a randomized controlled experiment in which n_t units are randomly sampled from a large population and assigned to a treatment group, and n_c units are randomly sampled from a large population and assigned to a control group. She then posits that $Y_i = \alpha + \beta D_i + \epsilon_i$, where Y_i is the observed outcome for unit i, D_i is an indicator variable for treatment assignment, α and β are model coefficients, and ϵ_i is an error term with zero expectation. Finally, she assumes that the ϵ_i are i.i.d. Why might the assumption that the ϵ_i are i.i.d. be invalid?

9. Suppose that the true data generating process is

$$Y_i = \alpha + \beta X_i + \gamma Z_i + e_i,$$

   where e_i is i.i.d. with $\mathrm{E}[e_i] = 0$. Using OLS regression on n observations, the researcher fits the parameters in the restricted model:

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

   Denote the OLS solution $(\hat{\alpha}^*, \hat{\beta}^*)$.

   (a) Define the fitted value for observation i as $\hat{Y}_i = \hat{\alpha}^* + \hat{\beta}^* X_i$. Show that $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$.

   (b) Which assumptions and conditions, if any, are necessary for the result in (a) to be true?

   (c) Is $\hat{\beta}^*$ generally unbiased for β? Show your reasoning by evaluating $\mathrm{E}[\hat{\beta}^*]$.

10. 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

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

   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?

2 Computer Assisted Portion

People often talk about “weak instruments” bias and associated issues. Some researchers suggest using OLS instead of 2SLS when the instrument is suspected to be weak. We want you to conduct some simulations to examine these claims in a simplified setting.

Suppose that there is some random vector (Y, D, Z) with p.m.f.,

$$f(Y, D, Z) = \begin{cases} 5/36 & : Y = 0, D = 0, Z = 0 \\ 15/36 & : Y = 0, D = 0, Z = 1 \\ 10/36 & : Y = 0, D = 1, Z = 1 \\ 1/36 & : Y = 1, D = 1, Z = 0 \\ 5/36 & : Y = 1, D = 1, Z = 1 \\ 0 & \qquad\qquad : otherwise. \end{cases}$$

The researcher’s inferential target is the Wald estimand,

$$\theta = \frac{\text{Cov}\left(Y, Z\right)}{\text{Cov}\left(D, Z\right)}.$$

In this setting, $\theta = 0$, but the researcher does not know θ. Consider the following two estimators of θ:

- the Wald IV estimator,
$$\widehat{\theta}_{IV} = \frac{\widehat{\text{Cov}}\left(Y, Z\right)}{\widehat{\text{Cov}}\left(D, Z\right)},$$
- the OLS estimator,
$$\widehat{\theta}_{OLS} = \frac{\widehat{\text{Cov}}\left(Y, D\right)}{\widehat{\text{Var}}\left(D\right)},$$

where $\widehat{\text{Cov}}\,(.)$ denotes the sample covariance and $\widehat{\text{Var}}\,(.)$ denotes the sample variance.

The researcher applies $\widehat{\theta}_{IV}$ and $\widehat{\theta}_{OLS}$ to n observations i.i.d. (Y, D, Z). We want you to assess the operating characteristics of $\widehat{\theta}_{IV}$ and $\widehat{\theta}_{OLS}$ using simulations. Use at least 1000 simulations to calculate your answers.

Important: You want you to calculate all of your estimates *conditional* on $n|\widehat{\text{Cov}}\left(D, Z\right)| > 1$. That is, if $|\widehat{\text{Cov}}\left(D, Z\right)| \leq 1/n$ in one of your simulated samples, you should discard the sample and draw a new one.

11. Calculate the (conditional) bias, standard error and root-mean-squared-error of $\widehat{\theta}_{IV}$ when $n = 100$.
12. Calculate the (conditional) bias, standard error and root-mean-squared-error of $\widehat{\theta}_{OLS}$ when $n = 100$.
13. Calculate the (conditional) bias, standard error and root-mean-squared-error of $\widehat{\theta}_{IV}$ when $n = 10000$.

14. Calculate the (conditional) bias, standard error and root-mean-squared-error of $\widehat{\theta}_{OLS}$ when $n = 10000$.

15. Discuss your findings. What do you conclude?

16. Why did we specify that you calculate the operating characteristics of your estimates conditional on $n|\widehat{\mathrm{Cov}}\,(D, Z)| > 1$? What would have happened had we not specified this? (You may answer using either theoretical arguments or simulations.)