Quantitative Empirical Methods Exam

Yale Department of Political Science, January 2019

The exam consists of three parts. You have seven hours to complete them.

Back up your assertions with formal derivations 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 (Pen-and-paper section) consists of five questions asking about your theoretical and technical understanding of the subject. There might be 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 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 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 Pen-and-paper section

1. Prove or disprove the following claims.

a) If X and Y are uncorrelated, then they are independent.

b) If X and Y are independent, then they are uncorrelated.

c) If X is mean independent of Y, then X and Y are uncorrelated.

d) If X and Y are uncorrelated, then X is mean independent of Y.

Recall that X is mean independent of Y when $\mathrm{E}[X \mid Y] = \mathrm{E}[X]$.

2. Let X be a Bernoulli (binary) random variable. You have a data set of 360 i.i.d. observations of X. There are 240 observations taking the value one, and 120 taking the value zero.

a) Parameterize the distribution of X with a single parameter.

b) In words, briefly explain how one would estimate the distribution of X using the plug-in principle.

c) In words, briefly explain how one would estimate the distribution of X using maximum likelihood.

d) Using either the plug-in principle or maximum likelihood, derive an estimator of the parameter you defined. What is the estimate using the provided data set?

e) Construct a 95% confidence interval for the parameter under a normal approximation.

3. You want to study whether of a positive, $W_i = 1$, or negative, $W_i = 0$, campaign message is best at persuading potential voters to vote for a particular candidate. You use a survey experiment with three positive messages, $M_i \in \{1, 2, 3\}$, and three negative messages, $M_i \in \{-1, -2, -3\}$, to investigate the question. Treatment is defined as $W_i = 1$ when $M_i \geq 0$ and $W_i = 0$ otherwise. The sample consists of n units. The outcome Y_i is whether the respondent would vote for the candidate delivering the message.

a) Assuming SUTVA for W_i with respect to Y_i (i.e., no-interference and no hidden versions of treatment), how many potential outcomes does each unit have? Note that "no hidden versions of treatment" is sometimes called "causal consistency" or "treatment invariance."

b) Assuming SUTVA for W_i with respect to Y_i, formally define an average treatment effect capturing the causal quantity of interest.

c) Assuming only no-interference, how many potential outcomes does each unit have?

d) Assuming only no-interference, formally define an average treatment effect capturing the causal quantity of interest.

The layout of the survey is completely determined by M_i, so a unit's response may be assumed to be fixed and unambiguous given M_i.

4. A paper includes a regression analysis with X and Z as the independent variables and Y as the dependent variable. The following table reports the results.

Table: Regression results

	(1)	(2)	(3)
X	26.96	26.87	30.49
	(2.26)	(2.41)	(3.79)
X^2		0.20	
		(0.31)	
Z	10.26	14.86	17.54
	(24.66)	(21.15)	(23.01)
$X \times Z$			−7.19
			(4.35)
Intercept	9.87	−13.08	11.54
	(21.12)	(21.90)	(21.22)
n	263	263	263
R^2	0.688	0.692	0.698

Note: Each column reports the coefficients from an ordinary least squares regression of Y on the variables indicated by the rows for three different specifications. The included variables for a specification are indicated by the rows with non-empty cells. Robust standard errors are given in parenthesis.

a) What is the estimate of $\mathrm{E}[Y \mid X = 3, Z = 1]$ using specification 1?

b) What is the estimate of $\mathrm{E}[Y \mid X = 3, Z = 1]$ using specification 2?

c) What is the estimate of $\mathrm{E}[Y \mid X = 3, Z = 1]$ using specification 3?

d) What is the marginal effect of X on Y when $X = 3$ and $Z = 1$ using specification 1?

e) What is the marginal effect of X on Y when $X = 3$ and $Z = 1$ using specification 2?

f) What is the marginal effect of X on Y when $X = 3$ and $Z = 1$ using specification 3?

Note that "marginal effect" refers to the estimated derivative of the regression function and not a causal effect.

You do not need to add, subtract, multiply, or divide for a satisfactory answer. For example, "$0.53 - 5 \times (1.44 + 3.03)$" would be an acceptable form of an answer.

5. Treatment $W_i \in \{0,1\}$ has been assigned to units by an unknown mechanism in a very large population. The outcome is binary, $Y_i \in \{0,1\}$, and SUTVA can safely be assumed for W_i with respect to Y_i. You draw a very large sample of observations from the population. The population is so large that the observations can be seen as i.i.d., and the sample is so large that the population distribution can be seen as known (i.e., sampling variability can be ignored).

You are interested in the average treatment effect $\tau = \mathrm{E}[Y_i(1) - Y_i(0)]$. You observe the following quantities: $\mathrm{E}[Y_i \mid W_i = 1] = 0.7$ and $\mathrm{E}[Y_i \mid W_i = 0] = 0.3$ and $\Pr(W_i = 1) = 0.5$.

a) Is τ identified? If so, prove it by deriving an expression of τ using the observed quantities. If not, derive bounds for the effect (i.e., lowest and highest possible values of τ consistent with the observed quantities).

b) You are told that treatment was randomly assigned independently and with equal probability to the units in the population. Does your answer to the previous question change with this information? If so, prove identification or derive new bounds. (The population is so large that variability from random treatment assignment is negligible and can be ignored.)

It turns out some units have missing outcomes. Let $R_i = 1$ denote that Y_i is observed, and $R_i = 0$ that it is not. Treatment W_i is observed no matter the value of R_i. You *cannot* rule out that missingness is affected by treatment, so it may be that $R_i(1) \neq R_i(0)$. SUTVA holds for W_i with respect to R_i, so $R_i = W_i R_i(1) + (1 - W_i)R_i(0)$.

The observed quantities are now $\mathrm{E}[Y_i \mid W_i = 1, R_i = 1] = 0.7$ and $\mathrm{E}[Y_i \mid W_i = 0, R_i = 1] = 0.3$ and $\Pr(W_i = 1) = 0.5$. You also observe that $\Pr(R_i = 1 | W_i = 1) = \Pr(R_i = 1 | W_i = 0) = r$ for some $0 < r < 1$.

c) Derive bounds on τ given $r = 0.75$. Treatment is still randomized.

d) Show that the width of the bounds (i.e., the difference between the upper and lower bounds) is equal to $2(1 - r)$ for any value of r. Treatment is still randomized.

e) One way to measure the information gained by an identification strategy in a particular empirical setting is the width of the identified bounds. In this case, zero width (i.e., point identification) is the most information, and a width of two (i.e., bounds $[-1, 1]$) is the least information. For which value of r does the setting in part **d** (i.e., known randomization and missingness) provide the same amount of information about τ as the setting in part **a** (i.e., unknown assignment mechanism but no missingness)?

The law of total expectation will be useful solving this, e.g.:

$$\mathrm{E}[Y_i(1)] = \Pr(R_i(1) = 1)\,\mathrm{E}[Y_i(1) \mid R_i(1) = 1] + \Pr(R_i(1) = 0)\,\mathrm{E}[Y_i(1) \mid R_i(1) = 0].$$

Also recall that $(A, B) \perp\!\!\!\perp C$ implies $\mathrm{E}[A \mid B] = \mathrm{E}[A \mid B, C]$.

2 Essay section

Please read the attached article: Jensen, Nathan M. 2003. "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment." International Organization 57:587–616.

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:

- What is the effect of regime type on FDI? How is this estimated? What assumptions underlie this estimate? Do these assumptions seem valid here?
- What are the scope conditions of these findings? Would the results generalize to other settings or populations? What additional analysis would you conduct to provide a more precise statement on the scope conditions of the results?

3 Computer assisted section

This question investigates estimation of treatment effects in sharp regression discontinuity designs. Treatment is deterministically assigned based on covariate X_i such that $W_i = 1$ when $X_i \geq 0$ and $W_i = 0$ otherwise. The population is very large, so the observations in your sample can be considered to be i.i.d. draws. SUTVA holds, so $Y_i(1)$ and $Y_i(0)$ are well-defined potential outcomes. The causal quantity of interest is the average treatment effect for units at the cut-off: $\tau_{\text{RD}} = \text{E}[Y_i(1) - Y_i(0) \,|\, X_i = 0]$. The conditional expectation functions of the potential outcomes are smooth (infinitely differentiable).

The distribution of X_i is uniform on the interval $[-10, 10]$. The observed outcome is distributed as:

$$Y_i = X_i^3 + 50X_i + 20W_iX_i + U_i,$$

where U_i is unobserved and normally distributed with mean 100 and standard deviation 250.

1. What is the conditional expectation function of Y_i given X_i and W_i?
2. Is the provided information enough to identify τ_{RD}? If so, derive the value of the effect.

Consider a class of estimators $\hat{\tau}(e)$ indexed by $e \in [0, 1]$ that estimates τ_{RD} as the difference in means between treated and control observations within a window around the cutoff given by $|X_i| \leq 10 \times 64^e n^{-e}$ for sample size n. In detail:

$$\hat{\tau}(e) = \frac{\sum_{i=1}^n W_i K_e(X_i) Y_i}{\sum_{i=1}^n W_i K_e(X_i)} - \frac{\sum_{i=1}^n (1 - W_i) K_e(X_i) Y_i}{\sum_{i=1}^n (1 - W_i) K_e(X_i)} \quad \text{where} \quad K_e(x) = \begin{cases} 1 & \text{if } |x| \leq 10 \left(\frac{64}{n}\right)^e, \\ 0 & \text{otherwise.} \end{cases}$$

3. Draw 2500 samples from the data generating process described above for each sample size $n \in \{2^6, 2^7, 2^8, \cdots, 2^{16}\} = \{64, \cdots, 65536\}$. For each sample, estimate the treatment effect using the estimators $\hat{\tau}(0)$, $\hat{\tau}(0.5)$ and $\hat{\tau}(1)$. Use the estimates to estimate the bias, standard error and root mean square error of the estimators. Report the results in three plots, one for each of the performance measure. Use log-scales for both axes, and let the x-axes denote sample size.
4. Based on the simulation results and your theoretical insights, which estimators (if any) of the three have the following properties:

   a) Unbiased in finite samples: $\text{E}[\hat{\tau}] = \tau_{\text{RD}}$,

   b) Asymptotically unbiased: $\text{E}[\hat{\tau}] \to \tau_{\text{RD}}$,

   c) Convergent: $\text{Var}(\hat{\tau}) \to 0$,

   d) Consistent: $\hat{\tau} \xrightarrow{p} \tau_{\text{RD}}$.

   Provide brief explanations, but no formal derivations are needed.
5. If the data generating process were unknown, which of the three estimators would you use? Motivate your answer.