Quantitative Empirical Methods Exam

Yale Department of Political Science, August 2020

You have 24 hours to complete the exam. The 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 (Theoretical section) consists of six shorter questions that can be answered with pen and paper. You are allowed to consult textbooks and other reference material, but the questions are written so that well-prepared students should be able to answer them without such references. *Advice*: There may 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.

Part 3 (Computer assisted section) will involve using statistical software to answer one longer exercise with several associated questions. A complete answer to Part 3 will include code and output, as well as your written answers. Most students will need to consult textbooks and other references to complete this part. *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 the whole exam, you are permitted access to any and all written materials, as well as unrestricted use of your own computer with access to the internet. The only restriction is that you may not interact with any person, online or otherwise.

Please turn in your answers as an email to colleen.amaro@yale.edu.

1 Theoretical section

1. Several properties of probabilities follows from the Kolmogorov axioms. Two important ones are *monotonicity* and the *complement rule*.

   (a) Formally state the monotonicity and complement rule properties.

   (b) Prove the two properties starting with the Kolmogorov axioms.

   (c) Consider modifying the Kolmogorov axioms by removing non-negativity. Do monotonicity and the complement rule still hold?

2. Suppose a scholar performs n independent hypothesis tests. Assume all null hypotheses hold, so we have n independent p-values all distributed uniformly on the unit interval $[0, 1]$. In terms of n, what is the probability of rejecting at least one of the tests at the 5% significance level ($p \leq 0.05$)?

3. Consider a population distribution (Y, D), where Y is real-valued and D is binary, taking values in $\{0, 1\}$. You draw a sample of n units from the population by first drawing $n_1 \leq n$ units from the subpopulation with $D = 1$. That is, the first n_1 units in the sample are drawn from the conditional distribution of (Y, D) given $D = 1$. Next, you draw $n_0 = n - n_1$ units from the subpopulation with $D = 0$. That is, the remaining n_0 units in the sample are drawn from the conditional distribution of (Y, D) given $D = 0$. This means that $D_i = 1$ for all $i \in \{1, 2, \ldots, n_1\}$ and $D_i = 0$ for all $i \in \{n_1 + 1, n_1 + 2, \ldots, n\}$ with probability one. The units are otherwise drawn i.i.d.

   Consider the estimator

$$\begin{aligned}\widehat{\tau} &= \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - D_i)Y_i \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i - \frac{1}{n_0}\sum_{i=n_1+1}^{n} Y_i. \end{aligned} \tag{1}$$

   (a) Derive the expectation of the estimator $\mathrm{E}[\widehat{\tau}]$ in terms of the conditional expectations of Y given D.

   (b) Derive the variance of the estimator $\mathrm{Var}[\widehat{\tau}]$ in terms of the conditional variances of Y given D. (Hint: the terms of the estimator are uncorrelated.)

   (c) Let $p = n_1/n$ be the share of the sample with $D_i = 1$. Suppose that n is so large that we can allow p to take on any value in the interval $[0, 1]$. In terms of the conditional variances of Y given D, what value of p minimizes the variance of the estimator? In words, explain your result and the underlying intuition.

(d) Consider when D is a treatment indicator and Y is an outcome potentially affected by the treatment. Provide sufficient conditions for this estimator to be unbiased for the average causal effect of D on Y. (Hint: Start by defining potential outcomes.)

4. A paper includes the following regression table, computed using ordinary least squares and robust standard errors. It reports the results from a regression conducted on 1000 survey respondents. The outcome is *Donations*, or respondent's donations to a senator's reelection campaign in US Dollars. We have two predictors:

   - $Ideology$: self-reported ideology, on a scale from -2 (Very Liberal) to 2 (Very Conservative).
   - $Income$: income, scaled as quantile in the US income distribution, on a scale from 0 to 1.

Dependent Variable: *Donations*			
	(1)	(2)	(3)
$Ideology$	1.725	0.835	3.401
	(0.455)	(0.640)	(1.494)
$Ideology^2$			1.317
			(0.517)
$Income$	0.140	1.121	0.587
	(0.363)	(0.854)	(0.913)
$Ideology \times$		1.067	0.365
$Income$		(0.568)	(0.635)
Intercept	2.539	1.466	1.851
	(0.775)	(1.054)	(1.145)
n	1000	1000	1000

The paper also includes the following summary statistics:

- $\overline{Ideology} = -1.144$.
- $\overline{Income} = 0.695$.

Consider the following inferential targets:

- $\theta_1 = \mathrm{E}[Donations | Ideology = 2, Income = 0.5]$
- $\theta_2 = \left. \frac{\partial \, \mathrm{E}[Donations | Ideology, Income]}{\partial Ideology} \right|_{Ideology=2, Income=0.5}$
- $\theta_3 = \mathrm{E}\left[\frac{\partial \, \mathrm{E}[Donations | Ideology, Income]}{\partial Ideology} \right]$

(a) In words, what are θ_1, θ_2 and θ_3?

(b) Under specification (1), what are the estimates of θ_1, θ_2 and θ_3?

(c) Under specification (1), compute a 95% normal approximation-based confidence interval for θ_2.

(d) Under specification (2), what are the estimates of θ_1, θ_2 and θ_3?

(e) Under specification (3), what are the estimates of θ_1, θ_2 and θ_3?

5. Suppose that you are interested in conducting an instrumental variables analysis on the joint distribution (Z, D, Y), where Z, D and Y are a real-valued instrumental variable, treatment variable, and outcome respectively.

   However, the measurement of Z is subject to measurement error, so you do not directly observe the true instrument Z. Instead, you observe draws from (Z^*, D, Y), where $Z^* = Z + U$ is the observed instrument, and U is some random measurement error.

   (a) Provide sufficient conditions for the IV estimator using the observed instrument Z^*,

   $$\widehat{\tau}_{\text{IV}^*} = \frac{\widehat{\text{Cov}}(Y, Z^*)}{\widehat{\text{Cov}}(D, Z^*)}, \tag{2}$$

   to converge to the same limit as the IV estimator using the true instrument Z,

   $$\widehat{\tau}_{\text{IV}} = \frac{\widehat{\text{Cov}}(Y, Z)}{\widehat{\text{Cov}}(D, Z)}. \tag{3}$$

   (These conditions should involve U, and you can assume that both $\widehat{\tau}_{\text{IV}^*}$ and $\widehat{\tau}_{\text{IV}}$ are convergent.)

   (b) What does your answer imply about measurement error in an instrumental variable analysis?

6. 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? Briefly summarize and critically assess the arguments made about these types of estimators. (Recommended length: 250-500 words.)

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: Mo, Cecilia Hyunjung and Katharine M. Conn. 2018. "When Do the Advantaged See the Disadvantages of Others? A Quasi-Experimental Study of National Service." *American Political Science Review* 112(4):721–741.

3 Computer assisted section

In this exercise, you will consider a high-dimensional data set. That is, a data set with more covariates than units, namely, $n = 100$ units and $p = 250$ covariates. Each unit in the population has a vector $\mathbf{X} = (X_1, X_2, \ldots, X_{250})$ consisting of 250 covariates, each independently drawn from a uniform distribution on the interval $[-1, 1]$. The dependent variable is generated as $Y = \alpha + \mathbf{X}\boldsymbol{\beta} + U$ for a real-valued intercept α, a vector of coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{250})$, and an error term U. The intercept is five: $\alpha = 5$. The coefficients are all equal to one: $\beta_1 = \beta_2 = \cdots = \beta_{250} = 1$. The error term is independently drawn from a standard normal distribution.

You will use the "least absolute shrinkage and selection operator" (LASSO) estimator to complete this exercise. The LASSO estimator has a tuning parameter, which is also called "regularization" or "penalty" parameter. This parameter is generally denoted with λ (i.e., the Greek letter lambda).

Some (but not all) of the following questions are best answered using a Monte Carlo simulation. For such a simulation, use at least 2,500 simulation iterations. The simulation should take no more than 10-15 minutes to run on a modern laptop computer, but you might first want to bug test your code with a fewer number of iterations.

1. What is the value of $\theta = \mathrm{E}[Y|X_1 = 1, X_2 = X_3 = \cdots = X_{250} = 0]$?
2. Explain why one generally cannot use the ordinary least squares (OLS) estimator to estimate θ.
3. Show and explain how one can use the LASSO estimator to estimate θ.
4. What are the bias, SE, and RMSE of the LASSO estimator of θ when the tuning parameter is $\lambda = 0.1$?
5. What are the bias, SE, and RMSE of the LASSO estimator of θ when the tuning parameter is $\lambda = 0.5$?
6. What are the bias, SE, and RMSE of the LASSO estimator of θ when the tuning parameter is $\lambda = 2$?
7. Explain why the properties of the LASSO estimator differ across questions 4–6.
8. Which of the three estimators in questions 4–6 would you use to estimate θ? Explain your motivation.