Quantitative Empirical Methods Exam

Yale Department of Political Science, January 2022

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. Suppose there is a population of n subjects, where exactly 55% of the subjects are Democrats and exactly 45% of the subjects are Republicans. You randomly sample one person from this population, and record whether or not this person was a Democrat or a Republican.

   (a) Formally represent this random generative process as a probability space.

   (b) Formally define a random variable X that takes on the value 0 if the sampled person is a Democrat, and 2000 if the sampled person is a Republican.

   (c) Find the PMF of X.

   (d) Find $\mathrm{E}[X]$.

2. Provide brief definitions of the following terms:

   (a) standard deviation

   (b) standard error

   (c) unbiasedness

   (d) confidence interval

   (e) p-value

   (f) statistical power

3. 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 $[0, 1]$.

   (a) In terms of n, what is the probability of rejecting at least one of the tests at the 5% significance level ($p \leq 0.05$)?

   (b) Explain how this example is related to p-hacking.

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. 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.)

6. Scholars have debated the practice of regression adjustment, or using regression models to adjust for covariate imbalance in randomized experiments. It is expected that you will use the internet to research this topic, and your answer should cite relevant readings and references.

   - Critically evaluate the arguments for and against the use of regression adjustment with OLS for randomized experiments. (recommended length: appx. 750 words)
   - Under what circumstances is regression adjustment with OLS approximately unbiased?
   - Under what circumstances does regression adjustment with OLS improve precision?

7. Consider a fuzzy regression discontinuity design with outcome Y, treatment D, assignment Z and continuous forcing variable X. Assume the cutpoint is at zero, so that

$$Z = \begin{cases} 1: & X \geq 0 \\ 0: & X < 0 \end{cases}.$$

   Suppose that you have a consistent estimator of

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

   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. Under these conditions, what population of units does this causal effect apply to?

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

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: Carnegie, A. and Marinov, N. (2017), Foreign Aid, Human Rights, and Democracy Promotion: Evidence from a Natural Experiment. *American Journal of Political Science*, 61: 671–683.

3 Computer assisted section

You want to investigate how income inequality in the US has changed during the COVID-19 pandemic. In particular, you will estimate the US Gini coefficient for the years 2018, 2019, 2020 and 2021. You will do this using a subsample from the Current Population Survey (CPS) contained in the file `income-data.csv`.

The data set contains three columns: `YEAR` is the year of the recorded data point; `SERIAL` is the household serial number; and `HHINCOME` is the household income for the corresponding year. You have five hundred observations from each year. You can safely assume that the data is independent both within and between the different years. You can also safely assume that the data is identically distributed according to each respective population distribution, which for the purposes of this exercise can be seen as continuous.[1]

The Gini coefficient is a measure of income inequality with a range from zero (minimum inequality) to one (maximum inequality). If you aren't already familiar with the Gini coefficient, it might be useful to take a quick read on Wikipedia. The formal definition of the coefficient is

$$G = \frac{1}{2\mu} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x)\, p(y)\, |x - y|\, dx\, dy,$$

where $p(\cdot)$ is the probability distribution function of the population income distribution and μ is the mean income in the population.

1. Construct a plug-in estimator for the Gini coefficient.
2. Estimate the US Gini coefficient for 2018, 2019, 2020 and 2021.
3. Estimate the standard errors of your four estimators.
4. Construct 95% confidence intervals for each Gini coefficient.
5. Present your results in a table.
6. Present your results in a plot.
7. You want to test whether the difference between the Gini coefficients for 2019 and 2020 is different from zero. Propose and justify a test. What is the p-value resulting from this test?
8. What conclusions (if any) do you draw from your investigation?

[1] This is real data from the Current Population Survey. However, the survey weights have been omitted to streamline the exercise, which means that the observations are not quite IID and that your estimator will be somewhat biased. You are expected to ignore this bias in this exercise.