Quantitative Empirical Methods Exam

Yale Department of Political Science, August 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 eight 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.

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 40% of the subjects are Democrats, exactly 25% are Independents, and exactly 35% of the subjects are Republicans. You randomly sample one person from this population, and record whether or not this person was a Democrat, Independent, 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, 100 if the sampled person is an Independent, and 400 if the sampled person is a Republican.

   (c) Find the PMF of X.

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

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

2. Assume that the conditional expectation function of Y given X and Z is:

$$\mathrm{E}[Y|X = x, Z = z] = 10 + x^3 + z + 4xz.$$

   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?

3. 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×		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:

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

Consider the following inferential targets:

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

(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?

4. Imagine a population is 95% vaccinated against a hypothetical virus and 51% of infected individuals have been vaccinated. A newspaper letter to the editor argues that this means the vaccine is ineffective at preventing infection. Is this letter to the editor correct? Explain why or why not.

5. Using election data, investigators make a study of the various factors influencing voting behavior. They estimate that the issue of inflation contributed 7 percentage points to the Republican vote in a certain election. However, the standard error for this estimate is 5 percentage points. The investigators conclude that "in fact, and contrary to widely held views, inflation has no impact on voting behavior." Does the conclusion follow from the statistical test? Answer yes or no, and explain briefly. Your answer should be about the standard error and not whether this is valid causal inference.

6. A recent study examined the income gains from migration. Indians need an Australian visa to migrate to Australia. A lottery is used to issue visas to applicants, and only lottery winners get a visa. However, some lottery winners do not end up migrating. A research group interested in the causal effects of migration on earnings presented a variety of estimators.

   (a) First, they present the difference between the mean weekly income of lottery winners that migrate and the mean income of lottery losers. One critic in the audience claims that since Indians self-select into migration, this comparison is misleading. The researchers argue, however, that the lottery generated random assignment, and thus the criticism is invalid. Who is correct and why? What is it that the researchers are estimating?

   (b) The researchers then present the difference between the mean income of all lottery winners and all lottery losers, and argue that this consistently estimates the average treatment effect of migration on earnings. The critic again claims that this is incorrect, now because not everyone who won the lottery migrated, and it is thus incorrect to attribute this difference to migration. Who is correct and why? What is it that the researchers are estimating?

   (c) Finally, the researchers show the difference in migration rates between the lottery winners and the lottery losers. They argue that this is a consistent estimate for the average treatment effect of the lottery on migration. The critic, once again, complains that this is incorrect because the lottery losers do not even have the option of migrating. Who is correct and why? What is it that the researchers are estimating?

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?

8. To measure the effect of exercise on the risk of heart disease, investigators compared the incidence of this disease for two large groups of London Transport Authority employees – drivers and conductors. The conductors got a lot more exercise as they walked around all day collecting fares. The age distributions for the two groups were very similar, and all the subjects had been on the same job for 10 years or more. The incidence of heart disease was substantially lower among the conductors, and the investigators concluded that exercise prevents heart disease. Other investigators were skeptical. They went back and found that London Transport Authority had issued uniforms to drivers and conductors at the time of hire; a record had been kept of the sizes.

   (a) Why does it matter that the age distributions of the two groups were similar?

   (b) Why does it matter that all the subjects had been on the job for 10 years or more?

   (c) Why might the second group of investigators have been skeptical?

   (d) What would you do with the sizes of the uniforms?

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:

> Kenneth Scheve and David Stasavage. “Democracy, War, and Wealth: Lessons from Two Centuries of Inheritance Taxation.” *American Political Science Review* 106:1 (2012), 81-102. `https://doi.org/10.1017/S0003055411000517`

3 Computer assisted section

You have designed and implemented a simple experiment with one treatment and one control group. The data can be found in `balance.csv`, and it contains an indicator for the treatment assignment (D) and five covariates (X_1, X_2, X_3, X_4, X_5). You had to rely on outside contractors to perform the treatment randomization, so you want to perform balance checks. Perform a randomization inference balance check that proceeds as follows:

(i) Use the Mahalonobis distance to operationalize the distance between any two data points as a function of the five covariates in the data:

$$\mathrm{d}(\overrightarrow{x_1}, \overrightarrow{x_2}) = \sqrt{(\overrightarrow{x_1} - \overrightarrow{x_2})^\top S^{-1}(\overrightarrow{x_1} - \overrightarrow{x_2})},$$

where $\overrightarrow{x_1}$ and $\overrightarrow{x_2}$ are each length-5 vectors for each unit and S is the covariance matrix.

(ii) Compute all pairwise distances between every treatment-control pair of data points. If there were 2 treated and 2 control units, that would be 4 pairwise comparisons.

(iii) Use the mean to summarize all pairwise distances between the observed treatment-control vectors.

(iv) Using randomization inference, compute a distribution of the mean difference assuming the sharp null of no imbalance. This should be done empirically across M iterations, where M is a number that you think is sufficient. Each iteration generates a different treatment assignment vector.

(v) Using this distribution under the sharp null, estimate the p-value for the observed mean difference.

Answer the following questions:

1. Implement the process above with the dataset, and report the p-value. Your implementation should be delivered in the form of a final function that takes as arguments: the treatment vector, the matrix of covariates, and the number of randomization iterations (M). You may **not** use wrapper packages such as `ri` or `ri2`.
2. Interpret the observed test-statistic, the sharp null, and the p-value in this context accurately in no more than three sentences.
3. A simpler balance check would have been to conduct a two-sample t-test of difference in means for each of the covariates, or conduct a F-test with a regression model regressing D on X_1, X_2, X_3, X_4, X_5. Explain what this simpler approach misses, and what the approach you implemented here improves upon.