Quantitative Empirical Methods Exam

Yale Department of Political Science, January 2024

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 short 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 1,000 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.

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. Provide brief definitions of the following terms:

   (a) variance

   (b) unbiasedness

   (c) consistency

   (d) confidence interval

   (e) p-value

   (f) statistical power

   (g) admissibility (of an estimator)

2. Consider the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Articulate sufficient conditions to interpret the OLS estimate of β_1 as a good approximation of a causal effect.

3. You run a randomized experiment on the effect of a postcard on voter turnout, such that you independently flip a coin with probability $0 < p < 1$ for each subject. Assume SUTVA. You estimate the following regression model using OLS: $Y_i = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 T_i S_i + u_i$ where Y_i indicates whether someone voted, T_i is the randomly assigned (binary) treatment indicator, S_i is an indicator variable for whether someone lives in a competitive battleground state or not.

   (a) Interpret each of the four β coefficients. When a coefficient has a causal interpretation, be sure to note it.

   (b) You want to test against the null hypothesis that the average treatment effect is identical for subjects living in competitive battleground states and those not living in competitive battleground states. How would you do that? Be specific.

4. You are conducting a randomized experiment to test whether TV ads increase voter turnout. You are able to conduct the experiment in up to 16 media markets and you are able to randomly assign up to 8 media markets to the treatment condition. For this experiment, TV ads can only be delivered to the entire media market.

   You have access to information for each eligible individual in each of the 16 media markets. For each individual, you know their media market, age, gender, race/ethnicity, and whether or not they voted in the prior two elections.

   After the election, you will receive an updated file to know whether or not each individual voted in the current election.

   (a) How would you randomly assign media markets to treatment or control? Be specific. Explain your reasoning.

(b) How would you analyze your proposed experiment? Be specific. Explain your reasoning.

5. Sometimes researchers analyzing a survey experiment drop subjects from their analysis because the subjects: (1) failed a pre-treatment attention check; (2) failed a post-treatment attention check; (3) took the survey too quickly (e.g., survey completion time was 3 standard deviations faster than the mean).

   (a) Suppose that the researcher is interested in the average treatment effect among subjects who are not dropped from the analysis. Which of these strategies (if any) are guaranteed to be unbiased?

   (b) Suppose now that the researcher is interested in the average treatment effect for all subjects who begin the survey experiment. Which of these strategies (if any) are guaranteed to be unbiased? (Assume there is no missingness; everyone who begins the survey experiment completes the survey experiment and answers every question in the survey.)

6. A die is rolled 8 times. When a die is rolled, each of the six faces is equally likely to come up. Find the chances of:

   (a) Getting 8 sixes.

   (b) Every roll showing 5 or less.

7. Suppose you have $n = 100$ observations of $p = 1000$ predictors $\mathbf{X} = (X_1, ..., X_{1000})$ and an outcome Y. Is it possible to estimate $\mathrm{E}[Y|\mathbf{X}]$? If so, discuss some ways in which you might estimate $\mathrm{E}[Y|\mathbf{X}]$. Under what circumstances would these approaches behave well?

8. 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 following article and 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 should be between 750 to **1000** words. We advise you to pay particular attention to the identification conditions (either explicit or implicit), the associated estimation strategy, possible threats to inference, statistical power, and generalizability. 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:

> Ferrer, Joshua, Igor Geyn, and Daniel Thompson. 2023. "How Partisan Is Local Election Administration?" *American Political Science Review.* `https://doi.org/10.1017/S0003055423000631`

Note: The article contains a appendix with relevant discussion and findings. We recommend you refer to some of those findings. We have provided you with a copy for your convenience.

3 Computer assisted section

Suppose you obtained a dataset characterized by the following three-way contingency table:

$X_i = 0$			$X_i = 1$		
	$Y_i = 0$	$Y_i = 1$		$Y_i = 0$	$Y_i = 1$
$D_i = 0$	23	42	$D_i = 0$	10	15
$D_i = 1$	33	91	$D_i = 1$	12	32

where the elements of the table refer to the number of observations that satisfy all three conditions. E.g., we have 15 observations with $Y_i = 1$, $D_i = 0$ and $X_i = 1$. Assume the data is sampled i.i.d. from a large population described by the joint distribution (Y, X, D).

Submit a script and written answers for the following questions. The script must be anonymous (it does not include file paths with your name on it).

1. Compute the conventional plug-in estimate of the parameter
$$\Delta = \mathrm{E}[Y|D = 1, X = 0] - \mathrm{E}[Y|D = 0, X = 0],$$
and construct a normal approximation-based 95% confidence interval for Δ.

2. Using a probit regression, estimate the coefficients $(\gamma_0, \gamma_1, \gamma_2)$ assuming that
$$\Pr[Y_i = 1|D_i, X_i] = \Phi(\gamma_0 + \gamma_1 D_i + \gamma_2 X_i)$$
where Φ is a standard Normal CDF.

3. Assuming that the probit model in Q2 is correct, compute the maximum likelihood estimate of
$$\Delta = \mathrm{E}[Y|D = 1, X = 0] - \mathrm{E}[Y|D = 0, X = 0],$$
and construct a normal approximation-based 95% confidence interval for Δ using the bootstrap.

4. Using a probit regression, estimate the coefficients $(\zeta_0, \zeta_1, \zeta_2, \zeta_3)$ assuming that
$$\Pr[Y_i = 1|D_i, X_i] = \Phi(\zeta_0 + \zeta_1 D_i + \zeta_2 X_i + \zeta_3 D_i X_i).$$

5. Assuming that the probit model in Q4 is correct, compute the maximum likelihood estimate of
$$\Delta = \mathrm{E}[Y|D = 1, X = 0] - \mathrm{E}[Y|D = 0, X = 0]$$
and construct a normal approximation-based 95% confidence interval for Δ using the bootstrap.

6. Using OLS with the specification

$$\mathrm{E}[Y_i | D_i, X_i] = \alpha_0 + \alpha_1 D_i + \alpha_2 X_i,$$

   estimate the coefficients $(\alpha_0, \alpha_1, \alpha_2)$.

7. Assuming that the linear specification in Q6 is correct, compute an estimate of

$$\Delta = \mathrm{E}[Y | D = 1, X = 0] - \mathrm{E}[Y | D = 0, X = 0]$$

   and construct a normal approximation-based 95% confidence interval for Δ using the bootstrap.

8. Using OLS with the interacted specification

$$\mathrm{E}[Y_i | D_i, X_i] = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i X_i,$$

   estimate $(\beta_0, \beta_1, \beta_2, \beta_3)$.

9. Assuming that the interacted specification in Q8 is correct, compute an estimate of

$$\Delta = \mathrm{E}[Y | D = 1, X = 0] - \mathrm{E}[Y | D = 0, X = 0]$$

   and construct a normal approximation-based 95% confidence interval for Δ using the bootstrap.

10. What conclusion, if any, can we draw about the parameter Δ from this analysis? Which specification do you prefer, and why?