Quantitative Empirical Methods Exam

Yale Department of Political Science, January 2023

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. Suppose an OLS regression produces a coefficient estimate of 1 with a standard error of 0.4. Rigorously explain the 95% Normal-approximation-based confidence interval associated with this regression, touching on two parts: What is the estimate of the confidence interval? If the confidence interval estimator used has proper coverage, what guarantees do these results provide?

2. Suppose a scholar performs n independent hypothesis tests. Answer the following about multiple hypothesis testing.

   (a) Assume the null hypothesis is true in all n cases, so we have n independent p-values all distributed uniformly on the unit interval $[0, 1]$. In terms of n, what is the probability that a test with a Type I error rate of 0.01 will reject at least one of the null hypotheses?

   (b) Now suppose that 1 percent of the n alternative hypotheses tested are actually true. We run a test with a Type I error rate of 0.01, but under an optimistic scenario, has a Type II error rate of 0. *Among the tests that turn out with a p-value of* $p < 0.01$, what proportion are false positives?

3. Is statistical significance "transitive" in the sense that if A is statistically significantly different from B and B is statistically significantly different from C, A is statistically significantly different from C? If Yes, explain why. If No, give a counterexample.

4. Assume the following data generating process for some outcome y:

$$y = \alpha + \beta x + \epsilon$$

   However, a researcher only observes measures of that contains error, denoted with an asterisk.

   (a) Consider a case in which the researcher observes the outcome estimated with error, y^*, where the error is defined as $e_y = y - y^*$. If the researcher estimates model $y^* = a + bx + u$ using OLS, under what assumption(s) will $\hat{b}$ be an unbiased estimate for $\hat{\beta}$? Will $\hat{\beta}$ be a consistent estimate of β under such assumption(s)? Show why or why not.

   (b) Now consider a case where the researcher measures x with error denoted as x^* where the error is denoted as $e_x = x - x^*$. The researcher (*with an accurate measure y, no longer* y^*) estimates the following model using OLS: $y = f + gx^* + v$. Assume that $E[x^* \times e_x] = 0$ and $E[e_x] = 0$. Will $\hat{g}$ be an unbiased estimate of β? Show why or why not.

5. Scholars have debated the practice of regression adjustment to improve the precision of treatment effect estimates 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.

   (a) Critically evaluate several arguments for and against the use of regression adjustment with OLS for randomized experiments. (recommended length: about 300 words)

   (b) Under what circumstances is regression adjustment with OLS approximately unbiased?

   (c) Under what circumstances does regression adjustment with OLS improve precision?

6. Suppose a researcher runs a fully randomized control trial in the field testing if a cash transfer to households in randomly selected villages improves saving rates a year out. This experiment is conducted across 100 villages in a particular region – 50 are assigned to treatment and 50 to control. After the randomization is complete and treatment households receive their transfer, an unexpected natural disaster decimates a fraction of the villages in the region where the experiment was taking place. You decide to drop households in these villages from your experiment and instead record outcomes for all households in the unaffected villages.

   (a) Under what assumptions will the difference in means between treatment and control households that you measure provide an unbiased estimate of the average causal effect of the cash transfer across all villages?

   (b) Under what assumptions will the difference in means between treatment and control households that you measure provide an unbiased estimate of an average causal effect of the cash transfer for some villages? To which villages does this apply?

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 should be between 500 to **1000** 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:

> Diamond, Rebecca, Tim McQuade, and Franklin Qian. “The effects of rent control expansion on tenants, landlords, and inequality: Evidence from San Francisco.” *American Economic Review* 109.9 (2019): 3365-94. `https://www.aeaweb.org/articles?id=10.1257%2Faer.20181289`

3 Computer assisted section

The US government wants to conduct an experiment to develop evidence-based best-practices for reducing the number of alcohol-impaired motor vehicle crash deaths. Counties in the US will be randomly assigned to a status quo control group or an educational campaign treatment group. In treatment counties, the US government will fund local non-profits to launch educational campaigns about the dangers of driving drunk.

Due to political constraints, each state must have 2 treated counties (50 states × 2 treated counties = 100; exclude DC) and will use an $\alpha = 0.05$ level test to reject the null hypothesis. Due to budget constraints, there can only be 100 treated counties in this experiment. This treatment costs $20 million per county to implement. Even though counties vary in their population, assume a uniform cost across counties. For this question, county population size does not matter. The US government values a life saved as being worth $10 million.

Given these parameters, can you design an experiment that will be sufficiently well-powered (e.g., power of 80%) to determine whether or not this program is cost-effective at saving a human life?

To assist you in conducting this power analysis, attached is historical data on the number of alcohol-impaired driving deaths by county from 2013-2020 from County Health Rankings. The variables include:

- `year`: The year of data this estimate represents
- `statecode`: The two-digit FIPS code for the state
- `countycode`: The three-digit FIPS code for the county
- `deaths`: The total number of alcohol-impaired motor vehicle crash deaths in that year and county

To answer this question, you should propose an experimental design, an analysis strategy, and the power associated with your hypothetical. You should try to maximize the power of your design, given the constraints described above and the data. Your submission should be both code and a *written explanation*, where you discuss your design, assumptions, and findings.

You do not need to search for any additional data to answer this question but you will need to make additional assumptions. All power calculations require assumptions. You should be sure to enumerate and briefly justify the assumptions you make. Because different people might make different but reasonable assumptions, there are many possible answers to this question. It is also ok to rely on some rounding of numbers. If so, just explain what you are doing.

One important assumption that we are giving you is constant treatment effects at the level of the county (i.e., $Y_i(1) = Y_i(0) + \tau$, where i indexes the county, τ represents the treatment effect, and Y a generic outcome). You do not need to justify this assumption.

While you may choose to answer the question differently, we suggest you answer this question by:

1. Determine the smallest possible treatment effect size that would be cost-effective for the US government.
2. Use the historical data to simulate a fake outcome variable.
3. Write code that conducts a random assignment and analysis, given the fake outcome variable.
4. Use a Monte Carlo simulation to approximate power.