Gov 2001: Final Exam

Spring 2025

May, 2025

Final Exam Instructions:

- This final exam is due on **May 5, 11:59 pm** Eastern time. Please upload a PDF of your solutions to **Gradescope**. When submitting, please match your responses with the questions.
- We will accept hand-written solutions, but we strongly advise graduate students to typeset your answers in LaTeX.
- This is a semi-closed book test. You are **NOT** allowed to: search the internet / AI for solutions or communicate amongst each other.
- You are allowed to utilize class materials (slides, section slides, pset solutions).
- Always use CGIS Knafel Zipcode **02138** as your seed for coding tasks.

1 OLS (30 pt)

Suppose you are studying the relationship between the number of hours a student studies (X_i) and their exam score (Y_i). Assume the following data-generating process for each student $i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where:

- $X_i \sim \text{Uniform}(0, 10)$, independent across i
- $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$, independent of X_i and independent across i
- $\beta_0 = 50$, $\beta_1 = 5$, and $\sigma^2 = 16$.

Answer the following:

(a) (6 points) Find the mean and variance of Y_i .

(b) (6 points) By the Law of Large Numbers (LLN), what happens to the sample mean $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ as $n \to \infty$?

(c) (6 points) By the Central Limit Theorem (CLT), approximate the distribution of $\sqrt{n}(\bar{Y}_n - \mathbb{E}[Y_i])$ as n becomes large.

(d) (6 points) Suppose you run an OLS regression of Y_i on X_i (with intercept). What are the probability limits (i.e., plim, convergence in probability) of the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ as $n \to \infty$?

(e) (6 points) Briefly explain why $\hat{\beta}_1$ is consistent under the given conditions (no need to formally prove, just check the necessary assumptions). What would happen if ε_i were correlated with X_i?

2 Publication Bias (30 pt)

Political science journals rarely publish statistically insignificant results. Does this publication bias lead to systematic bias in our understanding of political phenomena? Let's use simulations to find out. We will consider the effect of three different political phenomena—in each case, we are interested in the effect of X on Y. The following are the "true" models describing the relationship between each (X, Y) pair:

- $X_{1i} \sim N(0,1); \quad Y_{1i} = 2 + 0.1X_{1i} + u_{1i}; \quad u_{1i} \sim N(0,1)$
- $X_{2i} \sim N(0,1); \quad Y_{2i} = 2 + 5X_{2i} + u_{2i}; \quad u_{2i} \sim N(0,1)$
- $X_{3i} \sim N(0,1); \quad Y_{3i} = 2 + 0X_{3i} + u_{3i}; \quad u_{3i} \sim N(0,1)$

Thus, the only difference is the size of the true effect of X on Y. In the first model, X_1 has a very weak effect on Y_1; in the second, X_2 has a very strong effect on Y_2; and in the third, X_3 has no effect on Y_3.

1. (10pt) For each of the three phenomena, simulate 10,000 datasets of size 30 and calculate an OLS slope estimate for each. For each regression, record the difference between the estimated slope coefficient and the true value of β (i.e., the estimation error) and also record the p-value from the regression (you will use the p-value in the next part of this question). Plot the distribution (by histogram or density plot) of the *estimation error* among published articles for each of the three phenomena. Indicate the mean estimation error (an estimate for the bias) with a vertical line, and briefly interpret your plots in terms of bias.

   Hint: To get a p-value from the model, first you run a linear regression: `reg1 <- lm(y ~ X)`. Then the p-value is obtained by `summary(reg1)$coefficients[2,4]`.

2. (10pt) Now consider a journal editorial policy such that empirical research is not published unless the results are statistically significant, meaning that the p-value on the coefficient of interest is ≤ 0.05. Using your simulations from part (a), make a density plot for the difference between the estimated coefficients and the true value for all the *publishable* results under this policy. Plot each of the three phenomena separately. What percentage of studies are considered publishable for each (X, Y)

pairing? Do these studies correctly estimate the regression coefficient on average? Provide an intuitive explanation of the results. Should we be concerned by the findings?

3. (10pt) In the previous part, we assumed that only studies with significant results ended up getting published, but of course that is not realistic—null results are sometimes published. Let us assume instead that the journal is still willing to publish some null results, but significant results are much more likely to be published. Recreate your plots from part (b), but this time selecting articles for publication according to $\text{Bern}(p_i)$ where $p_i = 0.95$ if the study has significant results and $p_i = 0.05$ if the study does not. Separately for each phenomenon, sample a total of $n = 1000$ studies from the 10,000 you generated in part (a). Make sure to plot a vertical line at the mean of each of your plots to represent the bias. What do you notice? Under which phenomena should we be worried about this publishing practice?

   Hint: Recall weighted sampling in R from section zero, you can first generate a vector `prob_vec` with each unit's probability of being sampled, and then apply `sample(units, size = n, prob = prob_vec)`

3 Asymptotic Normality and Missing Data (40 pt)

You will conduct a simulation study to explore the impact of missing data on confidence intervals and estimator behavior. The true data-generating process is:

- $X_1 \sim \mathcal{N}(-4, 0.5)$
- $X_2 = 0.5X_1 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$
- $Y = 1 + 2 \cdot X_1 - 1 \cdot X_2 + \eta$, where $\eta \sim \mathcal{N}(0, 1)$

Missingness is introduced in X_1 according to:

$$\Pr(X_1 \text{ missing}) = \text{logit}^{-1}(X_2 + e)$$

where $e \sim N(1, 1)$, and

$$\text{logit}^{-1}(x) = \frac{1}{1 + e^{-x}}$$

1. (5pt) Explain why the missing data mechanism is Missing At Random (MAR).
2. (5pt) Write down the definition of asymptotic normality. What does it mean when a confidence interval has nominal coverage?
3. (10pt) Simulate 1000 datasets with $n = 500$ observations each. For each simulated dataset:
   - Estimate coefficients using (i) Oracle data (no missingness), (ii) Complete case analysis, and (iii) Multiple Imputation (5 imputations).

- Store point estimates and standard errors for β_{X_1} and β_{X_2}.
- You can use the R package `mice` for multiple imputation. Below is an example code, but feel free make necessary alterations or use other methods:

```
## Multiple Imputation
  library(mice)
  imp <- mice(dat_missing, m = 5, printFlag = FALSE)
  fit_mi <- with(imp, lm(Y ~ X1 + X2))
  pooled <- pool(fit_mi)
  summary_pool <- summary(pooled)
```

For each method and each coefficient, calculate the empirical coverage probability of confidence intervals at levels from 1% to 99%. Also, Write down *analytically* the general formula you use for any given level of confidence interval.

Hint: To generate missingness, you can first generate a `prob_vec` to store the probability of missing X_1 for each unit (you can check out the `plogis` function), and then generate a vector to indicate each unit's missingness by, for example, `miss_indicator <- rbinom(n, size = 1, prob = prob_vec)`.

4. (10pt) For each method, plot the empirical coverage curves (y-axis) against nominal confidence levels (x-axis, from 1% to 99%). Discuss:
   - Whether the estimators are approximately unbiased.
   - Whether the nominal confidence level matches the empirical coverage.
5. (10pt) Which method (Oracle, Complete Case, MI) appears most reliable under this missing data mechanism? Provide an explanation based on your findings. You can use more visualization to support your claims (for example, you can plot the distribution of simulated estimates under each method).