Gov 2001: Problem Set 8

Spring 2026

Instructions:

- The Problem set is due on **April 14, 11:59 PM Eastern Time**.
- Please upload a PDF of your solutions to Gradescope. Make sure to assign to each question all the pages with your work on that question.
- **Do not use AI assistants (ChatGPT, Claude, Copilot, etc.) on this problem set.** Work with each other instead. The struggle is where learning happens.
- Remember: 70% of your grade comes from in-class exams. Use problem sets to *learn*, not just to get answers.

Short Questions

1. Let $m(x) = \mathbb{E}[Y \mid X = x]$. Prove that $\mathbb{E}[XY] = \mathbb{E}[Xm(X)]$. (Hint: law of iterated expectations).

Long Questions

2. Let X and Y be random variables with CEF $\mathbb{E}[Y \mid X = x] = m(x)$. We want to use $l(x)$, a function of x, to approximate $m(x)$. We choose $l(x)$ by minimizing the mean squared prediction error:

$$\min_{l} \mathbb{E}\left[(Y - l(X))^2\right].$$

(a) Let $l_1(x) = \beta_1 x$. Find β_1 by solving the first order conditions.

(b) Let $l_2(x) = \alpha + \beta_2 x$. Find α and β_2 by solving the first order conditions.

Let $X \sim \text{Unif}(0, 1)$ and suppose the data are generated by

$$Y = m(X) + U, \qquad U \sim N(0, 1), \qquad U \perp X,$$

where

$$m(x) = 3 + 4x - 2x^2.$$

(c) Using your answers in Q1 and (a) and (b) to find a numerical expression of $l_1(x)$ and $l_2(x)$.

(d) (You can use AI for coding assistance in this question) Run a simulation with the following steps:

- Generate an iid random sample $(X_1, Y_1), ..., (X_n, Y_n)$ following the model with $n = 500$;
- Compute the empirical mean prediction errors

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - l_j(X_i))^2$$

for $j = 1, 2$.

- Repeat the simulation 10,000 times and compute the average mean prediction errors.

Compare your simulation results between $l_1(x)$ and $l_2(x)$. Which one has smaller mean prediction error? Why?

3. This question studies omitted variable bias and the Frisch–Waugh–Lovell (FWL) theorem.

Let X_2, W, and U be mutually independent standard normal random variables. Define

$$X_1 = X_2 + W, \qquad Y = 1 + 2X_1 + 3X_2 + U.$$

We define the following BLPs:

- Y on X_1 and X_2: $p(X_1, X_2) = a + bX_1 + cX_2$
- Y on X_1: $s(X_1) = \alpha + \beta X_1$
- Y on X_2: $q(X_2) = \alpha_Y + \delta_Y X_2$
- X_1 on X_2: $r(X_2) = \alpha_X + \delta_X X_2$

Finally, define the residualized variables

$$\tilde{Y} = Y - q(X_2), \qquad \tilde{X}_1 = X_1 - r(X_2).$$

(a) What are the values of a, b, and c?

(b) Derive $s(X_1)$ by computing α and β.

(c) Compare the coefficient on X_1 in (a) and (b). Why does omitting X_2 change the coefficient on X_1 here? Relate your answer to the correlation between X_1 and X_2.

(d) Derive $q(X_2)$ and $r(X_2)$. Then write $\tilde{Y}$ and $\tilde{X}_1$ as simple functions of U, W, and constants.

(e) Let $m(\tilde{X}_1) = \tilde{\alpha} + \tilde{\delta}\tilde{X}_1$ be the BLP of $\tilde{Y}$ on $\tilde{X}_1$. Compute $\tilde{\delta}$ and show that it is equal to the coefficient b from part (a). (Do not invoke the FWL theorem; directly compute the value and compare.)

(f) (You can use AI for coding assistance in this question.) Verify your answer in (e) by:

- generating a random sample of (Y, X_1, X_2) with $n = 500$;
- running a regression of Y on (X_1, X_2);
- running the auxiliary regressions of Y on X_2 and of X_1 on X_2, constructing $\tilde{Y}$ and $\tilde{X}_1$, and then regressing $\tilde{Y}$ on $\tilde{X}_1$;
- comparing the coefficient on X_1 in the full regression with the coefficient on $\tilde{X}_1$ in the residualized regression.