Problemset 1

Causal Inference and Machine Learning

LAURA MAYORAL
Instituto de Análisis Económico and Barcelona School of Economics

Winter 2026

INSTRUCTIONS:

(1) You can work individually or in groups, max., 3 people;
(2) If you work in groups, you can submit a group answer, clearly specifying the members of the group.
(3) Please submit via classroom.
(4) To access Wooldridge's datasets, follow the instructions given in this link: `http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html`
(5) Deadline: February 2nd.

QUESTIONS:

1. a) [Tip: read sections 10.2.2. and 10.2.3 in Wooldridge's book.] Describe the meaning of the strict exogeneity assumption and why it's needed in the estimation of fixed and random effects models.

b) Consider a model that explains sales in firm i as a function of its RD investments. We consider a *distributed lag model*, which is a model that includes current and past values of RD to explain current sales:

$$sales_{it} = \beta_1 RD_{it} + \beta_2 RD_{it-1} + c_i + u_{it}, \quad t = 1, \dots, T.$$

Discuss whether it's reasonable to expect that the strict exogeneity assumption will hold in this case.

2. Read Section 10.4 in Wooldridge's book. Consider the unobserved effects model for a randomly drawn cross-section observation i:

$$y_{it} = X_{it}\beta + c_i + u_{it}, \quad t = 1, \dots, T. \tag{1}$$

where X_{it} is a $1 \times K$ vector of regressors. Denote

$$X_i = (X_{i1}, \dots, X_{iT}), \quad u_i = (u_{i1}, \dots, u_{iT}). \tag{2}$$

Assume that the following conditions hold:

(i) $E[u_{it}|X_i, c_i] = 0, \quad t = 1, \dots, T$
(ii) $E[c_i|X_i] = E[c_i] = 0$

(iii) $E[u_i u_i' | X_i, c_i] = \sigma_u^2 I_T$ and $\text{Var}(c_i | X_i) = \text{Var}(c_i) = \sigma_c^2$.

a) Describe the meaning of assumptions (i), (ii) and (iii). Under these assumptions, does the OLS estimator have a causal interpretation?

b) Under assumptions (i)–(iii), what's the most efficient way of estimating model (1)? Add any other assumption you might need.

c) Now assume that assumptions (ii) and (iii) might not hold. Under assumption (i), does the OLS estimator have a causal interpretation?

d) Describe how you would estimate model (1) if assumption (ii) fails. Does this estimator have a causal interpretation under that assumption?

e) Describe how you would estimate the standard error of the estimator you suggested in (d) if (ii) and (iii) fail.

3. Problem 10.1. Wooldridge, page 291.

4. Problem 10.7. Wooldridge, page 294.

5. [Tip: Read section 21.4.3. in Cameron and Trivedi]

a) Describe the logic of the Hausman test to discriminate between random and fixed effects.

b) Describe the limitations of the non-robust version you computed in point 3) in the previous exercise.

c) Assess whether it's more reasonable in general to use FE versus RE models and why. Also, describe the potential problems of using FE estimators.

6. Problem 10.12. Wooldridge, page 296.

7. Answer the following questions.

a.) In the context of dynamic panel data model estimation, what is the Nickel bias? Explain clearly why the problem arises and whether it affects short and/or large panels similarly and why. Explain how the Arellano Bond estimator overcomes the problem.

b) Consider the original data in the Arellano and Bond (AB) paper (you can access it in STATA: webuse abdata). This is an unbalanced panel of annual data from 140 UK firms for 1976–1984. In their original paper, they modeled firms' employment n using a partial adjustment model to reflect the costs of hiring and firing, with two lags of employment. Estimate this model by OLS and using the within estimator. Compare the coefficient of the first lag of the dependent variable (employment) that you obtain in both models and discuss the direction of the bias that both

estimators are likely to have. (Hint: look at this document `http://fmwww.bc.edu/EC-C/S2014/823/EC823.S2014.nn05.slides.pdf`.

c) Using the same data, apply the AB estimator using the xtabond2 command. Discuss the output in detail, the difference between "GMM" and "IV" style instruments, as well as the meaning of the different options used in the document above.

d) Compare the value of the coefficient of the first lag of the dependent variable you've obtained in c): this value is in between the OLS and the within estimator. Discuss why the latter values can be interpreted as a lower and upper bound, respectively, for this coefficient.