Problemset 2

Causal Inference and Machine Learning

LAURA MAYORAL
Instituto de Análisis Económico and Barcelona School of Economics

Winter 2025

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`. You can access Hansen's datasets here.
(5) Deadline: February 21th.

QUESTIONS:

1. You want to estimate the probability of being unemployed as a function of a number of regressors, for instance age, level of education, etc. You have information for a group of N individuals over T=5 years (1 observation per year).

a) You would like to estimate this model including fixed effects to capture individual heterogeneous effects and for that you use a linear probability model. Describe how you could estimate it and discuss whether you would obtain consistent estimators and what would be the disadvantages of following this approach.

b) Explain why considering a nonlinear model can be useful in this case. Assuming you want to estimate a model with FE, explain the incidental parameter problem.

d) Explain what a sufficient statistic is and how it can help overcome the incidental parameter problem. What type of model you could estimate following this approach?

e) Use the KEANE dataset (Wooldridge) and construct a model to explain the probability of being employed (employ). Discuss whether the inclusion of FE can help to avoid the OVB. Employ FE and a nonlinear model to compute the estimates, justify your choice and provide an interpretation of the coefficients.

2. 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 and why. Explain how the Arellano Bond estimator overcomes the problem.

c) 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`.

d) 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.

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

3. Download the data (qreg0902.asc or qreg0902.data) for health expenditure from Cameron's website. (see mma04p2qreg.do for a description of this dataset and the variables in it). For each question, please clearly describe what you see.

a) Plot in the same graph two histograms for the log of household total expenditure and the log of household medical expenditure. Repeat using different values for the number of bins (one large value, one small value).

b) Use a Kernel estimate to estimate the density of the log of household medical expenditure. Choose the bandwidth using the plug-in estimator.

c) Repeat the same graph, trying different bandwidths (half and double of plug-in estimator, for instance).

d) Using the plug-in bandwidth estimator, estimate the density of these variables using three different kernels. Discuss the differences, if any.

e) Add a normal distribution to the estimated density (obtained using default values). Do the data look (log)normal?

f) Include the CI to the previous graph. You can try different ways of computing it. In particular, compute confidence intervals with values of the bandwidth different from the optimal one and describe the impact on the bias asymptotic behavior of choosing bandwidth values whose order of magnitude is larger or smaller than the optimal one.

4. Consider the data used in Exercise 19.9. in Hansen's book.

a) Plot a binned scatterplot of US gdp growth on the debt ratio. Repeat the same graph, this time controlling (in a careful way!) for the inflation rate. Discuss how binscatter and binsreg control for additional variables and which approach is more reliable and why.

b) Use Nadaraya-Watson to estimate the regression of gdp growth on the debt ratio. Plot with 95% confidence intervals. Discuss the role of the bandwidth in the estimation of this regression, and how it has been chosen. Try different bandwidths/kernels and observe if/how they affect the estimation.

c) Consider now a partially linear model (gdp growth on debt ratio and inflation) and estimate it using Robinson's semiparametric estimator. Estimate the model twice, considering first that the debt ratio enters nonparametrically and then the inflation rate. Dicuss the results.