Problemset 2

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) Each group should submit a zipped file containing:
   (a) A PDF document with the analytical/discussion parts of the solution, as well as all required Figures/Tables.
   (b) The code, which should run without any errors.
(3) Please submit the zipped file via Classroom. Only one submission per team.
(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. Cameron website: `https://cameron.econ.ucdavis.edu/mmabook/mmadata.html`. The `nlsw88` dataset is built into Stata. Load it with:

```
sysuse nlsw88, clear
```

(5) Deadline: February 17th.

QUESTIONS:

1. Non linear panel data models. 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 individual effects can help to avoid

the OVB. Employ a nonlinear model with individual effects to compute the estimates, justify your choice and provide an interpretation of the coefficients.

2. Histograms. Consider a random sample $X_1, \ldots, X_N$ drawn from some unknown continuous density f.

(1) Write down the formula for the histogram estimator $\hat{f}_{\text{HIST}}(x_0)$ centered at x_0 with bin width $2h$. Explain the role of h in this expression.
(2) Explain intuitively (no formal derivation needed) what happens to the histogram estimate as:
   (a) $h \to 0$ for fixed N,
   (b) $h \to \infty$ for fixed N,
   (c) $N \to \infty$ for fixed h.
(3) **Empirical.** Using the `nlsw88` dataset, construct histograms of the variable `wage` using 10, 20, and 50 bins. Comment on how the shape of the estimated density changes. Which number of bins seems most informative and why?

3. Kernel density Estimator.

(1) The kernel density estimator is given by

$$\hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right).$$

Explain the role of $K(\cdot)$ and h. How does this estimator relate to the histogram?
(2) State at least three properties that a kernel function $K(\cdot)$ must satisfy. Give two examples of commonly used kernels and write down their functional forms.
(3) Explain why the kernel density estimator can be seen as a "smoothed histogram." In particular, explain why using the uniform kernel $K(z) = \frac{1}{2}\mathbf{1}(|z| < 1)$ produces a "running histogram" rather than a standard histogram.
(4) Consider the Gaussian kernel, $K(z) = (2\pi)^{-1/2} \exp(-z^2/2)$, and the Epanechnikov kernel, $K(z) = \frac{3}{4}(1 - z^2)\mathbf{1}(|z| < 1)$. Which observations receive nonzero weight in each case? What are the practical implications of this difference?

4. Bandwidth selection and bias–variance trade-off.

(1) Define the Mean Squared Error (MSE) and the Mean Integrated Squared Error (MISE) of a kernel density estimator. Explain why MISE is a more useful criterion than MSE for evaluating the global performance of a density estimator.
(2) The optimal bandwidth satisfies $h^* = O(N^{-1/5})$. Explain intuitively why the optimal bandwidth shrinks as N grows.

(3) **Empirical.** Using the `nlsw88` data, estimate the kernel density of the variable `wage` using the Epanechnikov kernel with three bandwidths: the Stata default, half the default, and twice the default. Plot all three estimates on the same graph. Comment on the bias–variance trade-off visible in your plot.
(4) Using the plug-in bandwidth estimator, estimate the density of this variable using three different kernels. Discuss the differences, if any.
(5) On the same graph (or a new one), overlay a normal density with the same mean and standard deviation as `wage`. Does the wage distribution look normal? What features of the distribution does the kernel density estimator reveal that a parametric normal assumption would miss?

6. Binned scatter plots.

(1) Explain the steps involved in constructing a binned scatter plot. Why are binned scatter plots more informative than standard scatter plots when the sample size is large?
(2) The Stata command `binscatter` allows you to control for additional variables using the Frisch–Waugh–Lovell theorem. Explain how this works. Under what assumption on the conditional expectation is this approach valid? What can go wrong if this assumption is violated?
(3) **Empirical.** Using the `nlsw88` dataset:
[label=()]
   (a) Create a scatter plot of `wage` against `tenure`. Comment on what you can learn from it.
   (b) Create a binned scatter plot of `wage` against `tenure` using 20 quantile bins. Compare it with the scatter plot.
   (c) Now create the binned scatter plot controlling for `age`. Does the relationship change? Interpret.
   (d) Create the binned scatter plot separately by `race` (or `union` status). What do you observe?
(4) Cattaneo et al. (2024) propose an improved approach to binned scatter plots. Briefly describe two advantages of their approach over the traditional `binscatter` command.

6. Consider the model $y_i = m(x_i) + \varepsilon_i$, where $\mathbb{E}[\varepsilon|x] = 0$ and $m(\cdot)$ is unknown.

(1) Write down the Nadaraya–Watson (NW) kernel regression estimator $\hat{m}(x_0)$. Explain how it can be interpreted as a local weighted average.
(2) The NW estimator can also be obtained as the solution to a weighted least squares problem. Write down this optimization problem explicitly. *Hint:* think of fitting a constant locally.
(3) Explain the role of the bandwidth h in the NW estimator. What happens when h is very small? Very large? Relate this to the bias–variance trade-off.

(4) Describe the cross-validation procedure for selecting h in kernel regression. Why is it important to leave observations out when computing the cross-validation criterion? What problem does this prevent?
(5) **Empirical.** Using the `nlsw88` data, estimate the NW regression of `wage` on `tenure` using Stata's `lpoly` command (with `degree(0)`). Plot the estimate with confidence intervals. Experiment with at least two different bandwidths and comment on how the estimate changes.
Stata hint: `lpoly wage tenure, degree(0) ci`

7. Local linear regression.

(1) Write down the local linear regression estimator as the solution to a weighted least squares problem. How does it differ from the NW estimator?
(2) Name two advantages of local linear regression over the NW (local constant) estimator.
(3) The local linear estimator provides a direct estimate of $m'(x_0)$, the derivative of the conditional mean. Explain how. Why is this useful in applied work?
(4) **Empirical.** Using the `nlsw88` data:
[label=()]
(a) Estimate a local linear regression of `wage` on `tenure` using Stata's `npregress kernel` command (the default is local linear). Report the estimated average marginal effect of tenure on wages.
(b) Plot the estimated conditional mean function using `npgraph`.
(c) Compare your local linear estimate with: (a) the NW estimate from Problem 6, and (b) a simple OLS regression. Discuss the differences, paying special attention to the behavior at the endpoints.
Stata hint:

```
npregress kernel wage tenure
npgraph
```

8. Robinson's difference estimator. Consider the partially linear model:

$$y_i = x_i'\beta + \lambda(z_i) + u_i, \qquad \mathbb{E}[u_i \mid x_i, z_i] = 0,$$

where $\lambda(\cdot)$ is an unknown function.

(1) Explain why OLS estimation of β in the equation $y_i = x_i'\beta + (\lambda(z_i) + u_i)$ would be in general inconsistent (i.e., why we cannot simply ignore $\lambda(z_i)$ or treat it as part of the error term).
(2) Derive Robinson's difference estimator step by step:
(a) Take conditional expectations of both sides of the model, conditioning on z. What equation do you obtain?
(b) Subtract the equation from (i) from the original model. Write down the resulting equation.
(c) Explain how to estimate β from this differenced equation.

(3) Robinson's estimator of β is $\sqrt{N}$-consistent. Explain why this is remarkable—why doesn't the nonparametric estimation of the conditional expectations slow down the convergence rate?
(4) Describe how to estimate the nonparametric component $\lambda(z)$ once $\hat{\beta}$ has been obtained.

9. Robinson's estimator — empirical application. Using the `nlsw88` dataset, consider the following partially linear model:

$$\texttt{wage}_i = \beta_1 \cdot \texttt{union}_i + \beta_2 \cdot \texttt{married}_i + \lambda(\texttt{tenure}_i) + u_i.$$

Here, `union` and `married` enter the model linearly (as dummy variables), while the effect of `tenure` on wages is left unspecified.

(1) Before estimating the partially linear model, run a standard OLS regression of `wage` on `union`, `married`, and `tenure` (entering linearly). Report the estimated coefficients.
(2) Now estimate the partially linear model using Robinson's estimator.
Report the estimated coefficients on `union` and `married`. Compare them with the OLS estimates from part (a). Are the differences large? Interpret.
(3) The `semipar` command also produces a plot of the estimated $\hat{\lambda}(\texttt{tenure})$. Describe the shape of this function. Does the effect of tenure on wages appear to be linear? If not, how would you characterize the nonlinearity?
(4) Based on parts (a)–(c), discuss whether the linearity assumption for tenure in the OLS model seems reasonable. What are the practical implications of your findings?
(5) **Bonus.** Re-estimate the model using `ttl_exp` (total work experience) instead of `tenure` as the nonparametric variable. Does the nonparametric component look different? Comment.