Machine Learning for Economic Analysis
Problem Set 7

Jonas Lieber*

Due: 11:59pm Wed, March 13, 2024

Problem 1. *IV: Measurement error & Simultaneity*

1. *Consider the measurement error model discussed in class:*

$$Y = X\beta + U,$$

where we observe $(Y, \tilde{X}, \breve{X})$, *where*

$$\tilde{X} = X + V^1, \breve{X} = X + V^2,$$

and for $j \in \{1, 2\}$

   (a) $\mathbb{E}[V^j] = 0$,

   (b) $\mathbb{E}[XV^j] = 0$,

   (c) $\mathbb{E}[UV^j] = 0$,

   (d) $\mathbb{E}[V^1V^2] = 0$.

Verify exogeneity and relevance of $\breve{X}$ *as instrument for* $\tilde{X}$.

2. *Recall the following models for demand and supply considered in class*

$$Q^s(P) = \alpha^s + P\beta^s + Z\gamma + U^s,$$
$$Q^d(P) = \alpha^d + P\beta^d + U^d.$$

   (a) *Which sign do you expect for* β^s *and for* β^d? *Which sign do you expect for* $\beta^d - \beta^s$?

   (b) *Derive the equilibrium price.*

   (c) *Show that the equilibrium price is endogenous in the model for demand.*

   (d) *Verify exogeneity and relevance of* Z *as instrument for* P *in the demand model if if* $\gamma \neq 0$, $\mathbb{E}[ZU^d] = \mathbb{E}[ZU^s] = E[U^sU^d] = 0$ *and* $Var(Z) > 0$.

   (e) *Is* Z *also a valid instrument in the supply model?*

Problem 2. *Post-selection Inference & Double ML*

1. *Write a function to generate data for simulation.*

*Department of Economics, Yale University. jonas.lieber@yale.edu

(a) *The function should take n, the number of observations, $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}$ as inputs.*

(b) *The function should then create n i.i.d. data where*

i. $x_2 \sim \mathcal{N}(0,1)$, *the regressor we don't care about*

ii. $v \sim \mathcal{N}(0, 0.1)$ *is a noise parameter (the "exogenous part of x_1"), independent of x_2*

iii. $x_1 = x_2 + x_2^2\gamma_1 + x_2^5\gamma_2 + v$, *the regressor we do care about*

iv. $u \sim \mathcal{N}(0,1)$, *the "structural error",*

v. $y = x_1\beta_1 + x_2\beta_2 + x_2^2\beta_3 + sin(x_2)\beta_4 + u$

(c) *the function should return (y, x_1, x_2).*

2. *First consider the case $\gamma_1 = \gamma_2 = \beta_3 = \beta_4 = 0$. Write a function to calculate*

(a) *the estimated coefficient for β_1 of the "unrestricted" regression of y on x_1 and x_2 along with the standard error for this estimate*

(b) *the estimated coefficient for β_1 of the "restricted" regression of y on x_1 along with the standard error for this estimate*

(c) *the estimated coefficient for β_1 of the pretest regression discussed in class and by Leeb & Pötscher ("Model Selection and Inference: Facts and Fiction" 2005, Econometric Theory) along with an indicator for which model was selected and the standard error for the selected model*

3. *Consider $n = 10, 100, 1000$ and 10000 (so four values of n). For each n, suppose that $\beta_1 = 2$ and $\beta_2 = \frac{3}{\sqrt{n}}$ for each n. Then for each $s = 1, \ldots, S = 1000$,*

(a) *create a dataset using the function described in 1.*

(b) *compute the estimators using the function described in 2. and store the results.*

4. *To visualize the results,*

(a) *Create a table with mean and variance of each of the three ways to estimate the coefficients. Briefly interpret this result. Which phenomenon discussed in class do you observe?*

(b) *Report the fraction of instances (out of S) that the true model was selected. How does it vary with n? Interpret the result.*

(c) *For each of the three methods, report the fraction of instances (out of S) that the true coefficient is contained in the confidence interval $[\hat{\beta}_1 - 1.96\hat{\sigma}_1, \hat{\beta}_1 + 1.96\hat{\sigma}_1]$. Also report the average length of the confidence intervals. Interpret the results. Hint: Which coverage level would you expect?*

5. *Bonus (2 points): Plot the coverage probabilities for the pretest-model (with "oracle inference", as above) and the unrestricted regression for $n = 100$ as a function of β on a grid of beta (at least 20 values) that you have selected judiciously in a single plot.*

Problem 3. ***graduate students only*** *LASSO Theory*

For given $X \in \mathbb{R}^{n \times p}$*,* $\beta^* \in \mathbb{R}^p$ *and* $U \in \mathbb{R}^n$*, define* $Y = X\beta^* + U$*. For* $\lambda > 0$*, consider* $\hat{\beta}$*, defined as minimizer of* $f(\beta) = \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$*.*

1. *Show that for any* β

$$\|y - X\beta\|_2^2 = \|y - X\beta^*\|_2^2 + \|X(\beta^* - \beta)\|_2^2 + 2U^t X(\beta^* - \beta)$$

   Hint: Start on the left, add and subtract $X\beta^*$ *and use the definition of* U*.*

2. *Conclude that*

$$\left\|X(\hat{\beta} - \beta^*)\right\|_2^2 \leq 2U^t X(\hat{\beta} - \beta^*) + \lambda \left(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1\right)$$

   Hint: Use that $f(\hat{\beta}) \leq f(\beta^*)$*.*

3. *Where did we use this result in class?*