Lecture 11 – High-dimensional Microeconometric Models

April 29, 2016

Overview

Motivation

- **Machine Learning**: Methods usually tailored for prediction.
- In **Economics / Econometrics** both prediction (stock market, demand, ...) but also learning of relations / causal inference is of interest.
- Here: Focus on causal inference.
- Examples for causal inference: What is the effect of a job market programme on future job prospects? What is the effect of a price change?
- General: What is the effect of a certain treatment on a relevant outcome variable

Motivation

- Typical problem in Economics: potential endogeneity of the treatment.
- : Potential source: optimizing behaviour of the individuals with regard to the outcome and unobserved heterogeneity.
- Possible Solutions:
  - Instrumental Variable (IV) estimation
  - Selection of controls
- Additional challenge: high-dimensional setting with p even larger than n

Overview

Estimation and Inference with Many Instruments

Focus discussion on a simple IV model

$$\begin{aligned} y_i &= d_i\alpha + \varepsilon, && (1) \\ d_i &= g(z_i) + v_i, \text{(first stage)} && (2) \end{aligned}$$

with $\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} |z_i \sim \left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 \end{pmatrix}\right)$

- can have additional low-dimensional controls w_i entering both equations – assume these have been partialled out; also can have multiple endogenous variables; see references for details
- the main target is α, and g is the unspecified regression function $=$?optimal instrument?
- We have either
  - Many instruments. $x_i = z_i$, or
  - Many technical instruments. $x_i = P(z_i)$, e.g. polynomials, trigonometric terms.
- where where the number of instruments p is large, possibly much larger than n

Inference in the IV Model

- Assume approximate sparsity:

$$g(z_i) = E[d_i|z_i] = \underbrace{x_i'\beta_0}_{\text{sparse approximation}} + \underbrace{r_i}_{\text{approx error}}$$

  that is, optimal instrument is approximated by s (unknown) instruments, such that

$$s := ||\beta_0||_0 \ll n, \sqrt{1/n\sum_{i=1}^{n} r_i^2} \leq \sigma_v \sqrt{\frac{s}{n}}$$

- We shall find these "effective" instruments amongst x_i by Lasso and estimate the optimal instrument by Post-Lasso, $\hat{g}(z_i) = x_i'\hat{\beta}_{PL}$.
- Estimate α using the estimated optimal instrument via 2SLS

Example: Instrument Selection in Angrist Krueger Data

- y_i = wage
- d_i = education (endogenous)
- α = returns to schooling
- z_i = quarter of birth and controls (50 state of birth dummies and 7 year of birth dummies)
- $x_i = P(z_i)$, includes z_i and all interactions
- a very large list, $p = 1530$

Using few instruments (3 quarters of birth) or many instruments (1530) gives big standard errors. So it seems a good idea to use instrument selection to see if can improve.

AK Example

Estimator	Instruments	Schooling Coef	Rob Std Error
2SLS	(3 IVs) 3	.10	.020
2SLS	(All IVs) 1530	.10	.042
2SLS	(LASSO IVs) 12	.10	.014

Notes:

- About 12 constructed instruments contain nearly all information.
- Fuller's form of 2SLS is used due to robustness.
- The Lasso selection of instruments and standard errors are fully justified theoretically below

2SLS with Post-LASSO estimated Optimal IV

2SLS with Post-LASSO estimated Optimal IV

- In step one, estimate optimal instrument $\hat{g}(z_i) = x_i'\hat{\beta}$using Post-LASSO estimator.
- In step two, compute the 2SLS using optimal instrument as IV,

$$\hat{\alpha} = \left[1/n \sum_{i=1}^{n} (d_i \hat{g}(z_i)') \right]^{-1} 1/n \sum_{i=1}^{n} [\hat{g}(z_i) y_i]$$

IV Selection: Theoretical Justification

Theorem (2SLS with LASSO-selected IV)
Under practical regularity conditions, if the optimal instrument is sufficient sparse, namely $s^2 \log^2 p = o(n)$, and is strong, namely $|E[d_i g(z_i)]|$ is bounded away from zero, we have that

$$\sigma_n^{-1}\sqrt{n}(\hat{\alpha} - \alpha) \to_d N(0,1)$$

where σ_n^2 is the standard White?s robust formula for the variance of 2SLS. The estimator is semi-parametrically efficient under homoscedasticity.

- Ref: Belloni, Chen, Chernozhukov, and Hansen (Econometrica, 2012) for a general statement.
- A weak-instrument robust procedure is also available: the sup-score test
- Key point: "Selection mistakes" are asymptotically negligible due to "low-bias" property of the estimating equations, which we shall discuss later.

Example of IV: Eminent Domain

Estimate economic consequences of government take-over of property rights from individuals

- y_i = economic outcome in a region i, e.g. housing price index
- d_i = indicator of a property take-over decided in a court of law, by panels of 3 judges
- x_i = demographic characteristics of judges, that are randomly assigned to panels: education, political affiliations, age, experience etc.
- $f_i = x_i$ + various interactions of components of x_i ,
- a very large list $p = p(f_i) = 344$

Example continued

- Outcome is log of housing price index; endogenous variable is government take-over
- Can use 2 elementary instruments, suggested by real lawyers (Chen and Yeh, 2010)
- Can use all 344 instruments and select approximately the right set using LASSO.

Estimator	Instruments	Price Effect	Rob Std Error
2SLS	2	.07	.032
2SLS / LASSO IVs	4	.05	.017

Overview

Example: (Exogenous) Cross-Country Growth Regression.

- Relation between growth rate and initial per capita GDP, conditional on covariates, describing institutions and technological factors:

$$\underbrace{\text{GrowRate}}_{y_i} = \beta_0 + \underbrace{\alpha}_{\text{ATE}} \underbrace{\log(\text{GDP})}_{d_i} + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

where the model is exogenous,

$$E[\varepsilon_i | d_i, x_i] = 0.$$

- Test the convergence hypothesis – $\alpha < 0$ – poor countries catch up with richer countries, conditional on similar institutions etc. Prediction from the classical Solow growth model.
- In Barro-Lee data, we have p = 60 covariates, n = 90 observations. Need to do selection.

How to perform selection?

- (Don't do it!) Naive/Textbook selection
  1. Drop all x_{ij}s that have small coefficients, using model selection devices (classical such as t-tests or modern)
  2. Run OLS of yi on di and selected regressors.

  Does not work because fails to control omitted variable bias. (Leeb and Pötscher, 2009).
- We propose Double Selection approach:
  1. Select controls x_{ij}s that predict y_i .
  2. Select controls x_{ij}s that predict d_i .
  3. Run OLS of y_i on d_i and the union of controls selected in steps 1 and 2.
- The additional selection step controls the omitted variable bias.
- We find that the coefficient on lagged GDP is negative, and the confidence intervals exclude zero.

Method	effect	Std. Err.
Barro-Lee (Economic Reasoning)	-0.02	0.005
All Controls ($n = 90,\ p = 60$)	-0.02	0.031
Post-Naive Selection	-0.01	0.004
Post-Double-Selection	-0.03	0.011

- Double-Selection finds 8 controls, including trade-openness and several education variables.
- Our findings support the conclusions reached in Barro and Lee and Barro and Sala-i-Martin.
- Using all controls is very imprecise.
- Using naive selection gives a biased estimate for the speed of convergence.

TE in a PLM

Partially linear regression model (exogenous)

$$y_i = d_i\alpha_0 + g(z_i) + \xi_i, E[\xi_i|z_i, d_i] = 0,$$

- y_i is the outcome variable
- d_i is the policy/treatment variable whose impact is α_0
- z_i represents confounding factors on which we need to condition

For us the auxilliary equation will be important:

$$d_i = m(z_i) + v_i, E[v_i|z_i] = 0$$

- m summarizes the counfounding effect and creates omitted variable biases.

TE in a PLM

Use many control terms $x_i = P(z_i) \in \mathbb{R}^p$ to approximate g and m

$$y_i = d_i\alpha_0 +_i' \beta_{g0} + r_{gi} + \xi_i, d_i = x_i'\beta_{m0} + r_{mi} + v_i$$

- Many controls. $x_i = z_i$.
- Many technical controls. $x_i = P(z_i)$, e.g. polynomials, trigonometric terms.

Key assumption: g and m are approximately sparse

$$y_i = d_i\alpha_0 + x_i'\beta_{g0} + r_i + \xi_i, E[\xi_i|z_i, d_i] = 0,$$

Naive/Textbook Inference:

1. Select controls terms by running Lasso (or variants) of y_i on d_i and x_i
2. Estimate α_0 by least squares of y_i on d_i and selected controls, apply standard inference

However, this naive approach has caveats:

- Relies on perfect model selection and exact sparsity. Extremely unrealistic.
- Easily and badly breaks down both theoretically (Leeb and Pötscher, 2009) and practically.

(Post) Double Selection Method

To define the method, write the reduced form (substitute out d_i)

$$y_i = x_i'\bar{\beta}_0 + \bar{r}_i + \bar{\xi}_i, \quad (3)$$

$$d_i = x_i'\beta_{m0} + r_{mi} + v_i, \quad (4)$$

1. (Direct) Let $\hat{I}_1$ be controls selected by Lasso of y_i on x_i .
2. (Indirect) Let $\hat{I}_1$ be controls selected by Lasso of d_i on x_i .
3. (Final) Run least squares of y_i on d_i and union of selected controls:

$$(\tilde{\alpha}, \tilde{\beta}) = \arg\min_{\alpha,\beta} \left\{ 1/n \sum_{i=1}^{n} [(y_i - d_i\alpha - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \hat{I} = \hat{I}_1 \cup \hat{2}_1 \right\}.$$

The post-double-selection estimator.

- Belloni, Chernozhukov, Hansen (World Congress, 2010)
- Belloni, Chernozhukov, Hansen (ReStud, 2013)

Intuition

- The double selection method is robust to moderate selection mistakes.
- The Indirect Lasso step – the selection among the controls x_i that predict d_i – creates this robustness. It finds controls whose omission would lead to a "large" omitted variable bias, and includes them in the regression.
- In essence the procedure is a selection version of Frisch-Waugh procedure for estimating linear regression.

More Intuition

Think about omitted variables bias in case with one treatment (d) and one regressor (x):

$$y_i = \alpha d_i + \beta x_i + \xi_i, d_i = x_i + v_i$$

If we drop x_i , the short regression of y_i on d_i gives

$$\sqrt{n}(\hat{\alpha} - \alpha) = \text{good term} + \sqrt{n}(D'D/n)^{-1}(X'X/n)(\gamma\beta).$$

- the good term is asymptotically normal, and we want $\sqrt{n}\gamma\beta \to 0$.
- naive selection drops x_i if $\beta = O(\sqrt{\log n/n})$, but $\sqrt{n}\gamma\sqrt{\log n/n} \to \infty$
- double selection drops x_i only if both $\beta = O(\sqrt{\log n/n})$ and $\gamma = O(\sqrt{\log n/n})$, that is, if

$$\sqrt{n}\gamma\beta = O((\log n)/\sqrt{n}) \to 0.$$

Main Result

Theorem (Inference on a Coefficient in Regression)
Uniformly within a rich class of models, in which g and m admit a sparse approximation with $s^2 \log^2(p \vee n)/n \to 0$ and other practical conditions holding,

$$\sigma_n^{-1}\sqrt{n}(\hat{\alpha} - \alpha_0) \to_d N(0, 1)$$

σ_n^{-1} is Robinson's formula for variance of LS in a partially linear model. Under homoscedasticity, semi-parametrically efficient. Model selection mistakes are asymptotically negligible due to double selection.

Example: Effect of Abortion on Murder Rates

Estimate the consequences of abortion rates on crime, Donohue and Levitt (2001)

$$y_{it} = \alpha d_{it} + x_{it} + \xi_{it}$$

- y_{it} = change in crime-rate in state i between t and t - 1,
- d_{it} = change in the (lagged) abortion rate,
- x_{it} = controls for time-varying confounding state-level factors, including initial conditions and interactions of all these variables with trend and trend-squared
- p = 251, n = 576

Example continued

Double selection: 8 controls selected, including initial conditions and trends interacted with initial conditions

Estimator	Effect	Std. Err.
DS	-0.204	0.068
Post-Single Selection	-0.202	0.051
Post-Double-Selection	-0.166	0.216

Overview

Heterogenous Treatment Effects

- Here d_i is binary, indicating the receipt of the treatment,
- Drop partially linear structure; instead assume d_i is fully interacted with all other control variables:

$$y_i = d_i g(1, z_i) + (1 - d_i) g(0, z_i) + \xi_i, E[\xi_i | d_i, z_i] = 0$$

$$d_i = m(z_i) + u_i, E[u_i | z_i] = 0 (\text{as before})$$

- Target parameter. Average Treatment Effect:

$$\alpha_0 = E[g(1, z_i) - g(0, z_i)]$$

- Example. d_i = 401(k) eligibility, z_i = characteristics of the worker/firm, y_i = net savings or total wealth, α_0 = the average impact of 401(k) eligibility on savings.

Heterogenous Treatment Effects

An appropriate M_i is given by Hahn's (1998) efficient score

$$M_i(\alpha, g, m) = \left(\frac{d_i(y_i - g(1, z_i))}{m(z_i)} - \frac{(1 - d_i)(y_i - g(0, z_i))}{1 - m(z_i)} + g(1, z_i) - g(0, z_i) \right)$$

which is "immunized" against perturbations in g_0 and m_0:

$$\frac{\partial}{\partial g} E[M_i(\alpha_0, g, m_0)]|_{g=g_0} = 0, \frac{\partial}{\partial m} E[M_i(\alpha_0, g_0, m)]|_{m=m_0} = 0.$$

Hence the post-double selection estimator for α is given by

$$\tilde{\alpha} = 1/N \sum_{i=1}^{N} \left(\frac{d_i(y_i - \hat{g}(1, z_i))}{\hat{m}(z_i)} - \frac{(1 - \hat{d}_i)(y_i - \hat{g}(0, z_i))}{1 - \hat{m}(z_i)} + \hat{g}(1, z_i) - \hat{g}(0, z_i) \right)$$

where we estimate g and m via post- selection (Post-Lasso) methods.

Heterogenous Treatment Effects

Theorem (Inference on ATE)
Uniformly within a rich class of models, in which g and m admit a sparse approximation with $s^2 \log^2(p \vee n)/n \to 0$ and other practical conditions holding,

$$\sigma_n^{-1}\sqrt{n}(\tilde{\alpha} - \alpha_0) \to_d N(0,1)$$

where $\sigma_n^{-1} = E[M_i^2(\alpha_0, g_0, m_0)]$. Moreover, $\tilde{\alpha}$ is semi-parametrically efficient for α_0.

- Model selection mistakes are asymptotically negligible due to the use of "immunizing" moment equations.
- Ref. Belloni, Chernozhukov, Hansen, Inference on TE after selection amongst high-dimensional controls (Restud, 2013).