High-Dimensional Methods and Inference on Structural and Treatment Effects

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Machine Learning

Outline

1. Belloni et al.
2. Recap: Lasso
3. Causal Inference
   - Selecting IVs
   - Selecting Controls
4. Empirical Examples
   - Estimating the Impact of Eminent Domain on House Prices
   - Estimating the Effect of Legalized Abortion on Crime
   - Estimating the Effect of Institutions on Output

Belloni et al. (2014)

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2), 29-50.

- Many statistical methods are good at prediction but not suitable for making **inference** about model parameters when facing high-dimensional data.
- **Research Question**
  How innovations in “data mining" can be adapted and modified to provide high-quality inference about model parameters.

Recap: lasso

Suppose we are interested in forecasting outcome y_i with controls w_i according to the model

$$y_i = g(w_i) + \zeta_i$$

where $E[\zeta_i | w_i] = 0$, and $\forall i = 1, \cdots, n$ are independent observations.

- To **avoid overfitting** and produce useful **out-of-sample forecasts**, we generally **restrict or regularize** $g(\cdot)$.
- Here, we focus on regularization that treats $g(w_i)$ as a high-dimensional, approximately linear model. Assume

$$g(w_i) = \sum_{j=1}^{p} \beta_j x_{ij} + r_{pi}.$$

- $x_i = (x_{i1}, \ldots, x_{ip})'$ can be:
  - elementary regressors w_i, or
  - transformations of w_i in series modeling, allowing $p > n$.
- r_{pi} represents an approximation error, assumed to be minor compared to sampling error.
- The model uses an **approximately sparse** structure:
  - Only $s \ll n$ of the x_{ij} variables have nonzero coefficients β_j, with allowable approximation error r_{pi}.
  - **Lasso** is preferred for parameter estimation in these models.

- We use a variant of the lasso estimator Belloni et al. (2012)

$$\hat{\beta} = \arg\min_{b} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} b_j)^2 + \lambda \sum_{j=1}^{p} |b_j| \gamma_j, \text{ where}$$

  - $\lambda > 0$ is“penalty level", which controls the degree of penalization;
  - γ_j are “penalty loadings", which can insure basic equivariance of coefficient estimates to rescaling of x_{ij}, and can address hetero- skedasticity, clustering, and non-normality in model errors.

- Lasso advantages:
  - Promotes sparsity with many coefficients set to zero.
  - Solvable via efficient computational methods due to its convex nature.
  - Handles approximation errors, heteroskedasticity, clustering, fixed effects, and non-normality.
  - Nonzero coefficients are biased towards zero, but this can be corrected with a post-lasso estimator.

Caution is advised when using lasso results for inferential conclusions on model parameters.

When Goal is Causal Inference

- Regularization and model selection are geared towards forecasting, not parameter inference.
- **Model selection errors may occur.**
  - Variables with minor effects are often overlooked, leading to errors.
  - Neglecting these variables skews the estimation and inference of the selected variables.
- Inference procedures that are robust to model selection mistakes:
  1. Concentrate on a limited set of parameters, leaving regularization for the nuisance components.
  2. Apply estimating equations that are robust to minor changes in the nuisance components to accurately estimate the main parameters.

Outline

1. Belloni et al.
2. Recap: Lasso
3. Causal Inference
   - Selecting IVs
   - Selecting Controls
4. Empirical Examples
   - Estimating the Impact of Eminent Domain on House Prices
   - Estimating the Effect of Legalized Abortion on Crime
   - Estimating the Effect of Institutions on Output

Selection among IVs

Consider the linear instrumental variables model with potentially many instruments:

$$y_i = \alpha d_i + \epsilon_i$$
$$d_i = z_i' \Pi + r_i + v_i, \text{ where}$$

- $E[\epsilon_i | z_i] = E[v_i | z_i, r_i] = 0$ but $E[\epsilon_i v_i] \neq 0$, leading to endogeneity.
- d_i is a scalar endogenous variable of interest.
- z_i is a p-dimensional vector of instruments ($p \gg n$ is allowed).
- r_i is an approximation error.

To estimate and infer about α, select a few instruments from z_i for two-stage least squares estimation.

- Variable selection is confined to the first-stage's predictive equation.
- Omitting a valid instrument with a minor coefficient doesn't notably affect the estimate of α, provided that other instruments with substantial coefficients are used.
- Second-stage estimates are robust against errors from omitting instruments with small coefficients.

Outline

1. Belloni et al.
2. Recap: Lasso
3. Causal Inference
   - Selecting IVs
   - Selecting Controls
4. Empirical Examples
   - Estimating the Impact of Eminent Domain on House Prices
   - Estimating the Effect of Legalized Abortion on Crime
   - Estimating the Effect of Institutions on Output

Selection among Controls

Consider a linear model where a treatment variable, d_i, is taken as exogenous after conditioning on control variables:

$$y_i = \alpha d_i + x_i'\theta_y + r_{yi} + \zeta_i, \text{ where}$$

- $E[\zeta_i | d_i, x_i, r_{yi}] = 0$.
- x_i is a p-dimensional vector of controls ($p \gg n$ is allowed).
- r_{yi} is an approximation error.
- α is the parameter of interest, the effect of the treatment on the outcome.

Naive approach: Apply lasso to select control variables, keeping the treatment variable fixed by not penalizing α.

- Variables highly correlated with the treatment are dropped, as they add little predictive value when the treatment is included.
- This method is referred to as the naive **post-lasso**.

Two key issues with the naive approach:

1. Neglects the relationship between treatment and controls. Introduce a reduced form relation:

$$d_i = x_i'\theta_d + r_{di} + v_i, \text{ where } E[v_i|x_i, r_{di}] = 0.$$

2. Focuses on learning treatment effects, not forecasting. Transform $y_i = \alpha d_i + x_i'\theta_y + r_{yi} + \zeta_i$ into a structural equation:

$$y_i = x_i'(\alpha\theta_d + \theta_y) + (\alpha r_{di} + r_{yi}) + (\alpha v_i + \zeta_i) = x_i'\Pi + r_{ci} + \epsilon_i,$$
$$d_i = x_i'\theta_d + r_{di} + v_i$$

$E[\epsilon_i|x_i, r_{ci}] = 0$, r_{ci} represents composite errors.

- Equations predict relationships.
- Using one equation risks significant omitted-variables bias.

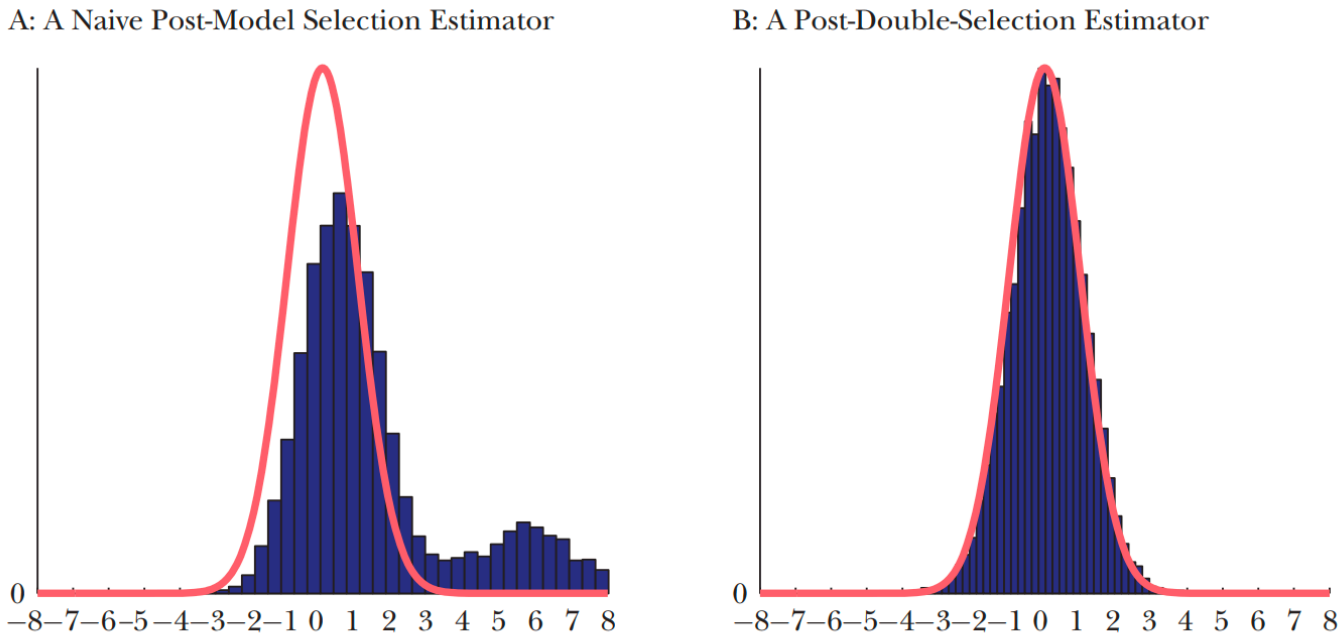
We apply variable selection methods to both equations and then use all of the selected controls (the union) to estimate α. Steps are as follows:

1. Use lasso on $d_i = x_i'\theta_d + r_{di} + v_i$ to select controls that are useful for predicting d_i, say x_{di}
2. Use lasso on $y_i = x_i'\Pi + r_{ci} + \epsilon_i$ to select controls that are useful for predicting y_i, say x_{yi}.
3. Estimate α using the OLS regression of y_i against d_i and controls in $x_{di} \cup x_{yi}$.

This approach is called the **double lasso**.

Another way that avoids the bias of the naive post-lasso estimator is the **partialling out** approach (Chernozhukov et al. (2016)):
estimate residuals ϵ_i and v_i and then regresses the estimates of ϵ_i on the estimates of v_i to construct an estimator of α.

Figure: Naive vs. Double Selection (Bellonie et al. (2014), 38.)



Source: Belloni, Chernozhukov, and Hansen (forthcoming).
Notes: The left panel shows the sampling distribution of the estimator of α based on the first naive procedure described in this section: applying LASSO to the equation $y_i = d_i + x_i' \theta_y + r_{yi} + \zeta_i$ while forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. The right panel shows the sampling distribution of the "double selection" estimator (see text for details) as in Belloni, Chernozhukov, and Hansen (forthcoming). The distributions are given for centered and studentized quantities.

Outline

1. Belloni et al.
2. Recap: Lasso
3. Causal Inference
   - Selecting IVs
   - Selecting Controls
4. Empirical Examples
   - Estimating the Impact of Eminent Domain on House Prices
   - Estimating the Effect of Legalized Abortion on Crime
   - Estimating the Effect of Institutions on Output

Estimating the Impact of Eminent Domain on House Prices

Chen, D. L., & Yeh, S. (2012). Growth Under the Shadow of Expropriation? The Economic Impacts of Eminent Domain.

- **Background**
  Federal court rulings that a government seizure was unlawful (pro-plaintiff rulings) uphold individual property rights and can make future exercise of eminent domain more difficult due to the legal system's structure in the US.
- **Research Question**
  What is the effect of eminent domain on house prices?

  There is endogeneity, e.g., a taking may be less likely if real estate prices are low and sellers are eager to unload property.

Identification Strategy

- Judges are randomly assigned to three-judge panels to decide appellate cases. The judges' identities and demographics are randomly assigned based on the distribution of characteristics of federal circuit court judges in a given circuit-year.
- Characteristics of judges serving on federal appellate panels can only be related to property prices through the judges' decisions (exclusion restriction).

Model

Uncover the effect of takings law by estimating models of the form

$$\log(Case - Shiller_{ct}) = \alpha \cdot TakingsLaw_{ct} + \beta_c + \beta_t + \gamma_c t + W'_{ct}\delta + \epsilon_{ct}$$

using the characteristics of judges actually assigned to cases as instruments for $TakingsLaw_{ct}$.

- $Case - Shiller_{ct}$: average of the Case-Shiller home price index within circuit court c at time t.
- $TakingsLaw_{ct}$: number of proplaintiff appellate takings decisions in federal circuit court c at time t.
- W_{ct}: exogenous variables.
- β_c; β_t; $\gamma_c t$: circuit-specific effects; time-specific effects; circuit-specific time trends.
- α: effect of an additional decision upholding individual property rights on a measure of property prices.

Instrument Selection I

Judges' Characteristics as Instruments:

- Characteristics satisfy the instrumental variables **exclusion restriction**.
- Infeasible to use all combinations of characteristics as instruments due to dimensionality.

Selection Process:

1. Perform dimension reduction to select characteristics with strong signals about judge preferences regarding government vs. individual property rights.
2. With 147 instruments, estimate the first-stage relationship using lasso, identifying one significant instrument: the *square of the number of panels with one or more members with JD from a public university.*
3. Conduct Two-Stage Least Squares (TSLS) using the selected instrument:
   - **First stage:** Coefficient = 0.4495, Standard Error (SE) = 0.0511 (strong IV).
   - **Second stage:** Coefficient = 0.0648, SE = 0.0240 (statistically significant).

Instrument Selection II

Comparison with Intuitive Instrument:

- *Intuitive* IV: *Number of judicial panels with one or more Democrats.*
- Found to be too weak for meaningful analysis.

Conclusion:

- Formal variable selection method leads to a stronger first-stage relationship and more precise second-stage estimate.
- High-dimensional techniques can complement intuition in instrument selection and enhance the ability to draw conclusions from data.

Outline

1. Belloni et al.
2. Recap: Lasso
3. Causal Inference
   - Selecting IVs
   - Selecting Controls
4. Empirical Examples
   - Estimating the Impact of Eminent Domain on House Prices
   - Estimating the Effect of Legalized Abortion on Crime
   - Estimating the Effect of Institutions on Output

Estimating the Effect of Legalized Abortion on Crime

Donohue III, J. J., & Levitt, S. D. (2001). The impact of legalized abortion on crime. *The Quarterly Journal of Economics*, 116(2), 379-420.

- **Background**
  Higher rates of abortion in the years around 1970, as legal restrictions on abortion were eased in a number of states, are associated with lower rates of crime two decades later.
- **Research Question**
  What is the effect of legalized abortion on crime?

  State-level abortion rates during the earlier time period were not randomly assigned; certain factors may be associated with both state-level abortion rates and state-level crime rates. Failing to control for these factors will then lead to OVB in the estimated abortion effect.

Difference-in-Differences

To address these potential confounding factors, estimate a DiD model (1985 - 1997, 600 obs.)

$$y_{cit} = \alpha_c a_{cit} + w_{it}'\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit}.$$

- y_{cit}: crime-rate for crime c (categorized between violent, property, and murder) in state i, time t.
- a_{cit}: abortion rate relevant for type of crime c.
- w_{it}: a set of variables to control for time-varying confounding state-level factors.
- δ_{ci}: time-invariant state-specific effects.
- γ_{ct}: time-specific effects that control for national aggregate trends.

[Result: 1$^{\text{st}}$ row in table]

Why Double-selection?

To obtain valid estimates of the causal effect of abortion on crime rates, it is important to capture time-varying state-specific factors that are correlated to both abortion and crime rates.

- Including a set of state-specific linear time trends may be inappropriate because
  - It introduces many additional variables.
  - Assuming trends to be linear may be questionable.
- Allowing for nonlinear trends interacted with observed state-specific characteristics and using variable selection methods can be better.
  - Accommodating a flexible trend that offers a sensible model of the evolution of abortion or crime rates over 12 years.
  - Using the **double-selection procedure** (double lasso) to identify potentially important confounding variables.

Double-selection Models

The equations are as follows:

$$\Delta y_{cit} = \alpha_c \Delta a_{cit} + z'_{cit}\beta_c + \tilde{\gamma}_{ct} + \Delta\epsilon_{cit}$$
$$\Delta a_{cit} = z'_{cit}\Pi_c + \tilde{\kappa}_{ct} + \Delta v_{cit}$$

- $\Delta y_{cit} = y_{cit} - y_{cit-1}$, same for $\Delta a_{cit}, \Delta\epsilon_{cit}, \Delta v_{cit}$.
- $\tilde{\gamma}_{ct}, \tilde{\kappa}_{ct}$: time effects.
- z_{cit}: 284 controls. Including state-specific time-varying variables, interactions of the variables with t and t^2, and the main effects t and t^2. (corresponds to a cubic trend of crime and abortion rate)

Why Not All Controls?

Theoretically, we can include all controls since # controls $<$ # obs. But the estimated abortion effects are **imprecise**.[2^{nd} row in table]

We can use the double-selection method [see 3^{rd} row in table]:

1. Select variables from z_{cit} that are useful for predicting the change in crime rate Δy_{cit} and Δa_{cit}.
2. Use the union of the set of selected variables, including time effects, as controls in a final ordinary least squares regression of Δy_{cit} on Δa_{cit}.

Model

Figure: Effect of Abortion on Crime (Donohue and Levitt (2001))

	Type of crime					
	Violent		*Property*		*Murder*	
Estimator	*Effect*	*Std. error*	*Effect*	*Std. error*	*Effect*	*Std. error*
First-difference	−.157	.034	−.106	.021	−.218	.068
All controls	.071	.284	−.161	.106	−1.327	.932
Double selection	−.171	.117	−.061	.057	−.189	.177

- Using a small set of intuitively selected controls: increases in abortion have a strong negative effect on crime rates.
- Using formal variable selection: fail to reject the hypothesis that abortion is unrelated to crime rates.
- It implies that 1^{st} row is not robust to the presence of fairly parsimonious nonlinear trends.

Outline

1. Belloni et al.
2. Recap: Lasso
3. Causal Inference
   - Selecting IVs
   - Selecting Controls
4. Empirical Examples
   - Estimating the Impact of Eminent Domain on House Prices
   - Estimating the Effect of Legalized Abortion on Crime
   - Estimating the Effect of Institutions on Output

Estimating the Effect of Institutions on Output

Study: Acemoglu, Johnson, & Robinson (2001) study the effect of institutions on output?

- **IV:** Simultaneity issue addressed using *early European settlers' mortality rates* as an IV for institutional quality.
- **Validity:** Better institutions were established in areas with lower initial settler mortality, showing high persistence.
- **Exclusion Restriction:** GDP's link to past mortality is primarily through institutional quality.
- **Identification Strategy:** Mortality risk serves as a valid IV when accounting for geography, which influences both institutional development and GDP.

Two-stage Least Squares

Normal IV-2SLS is

$$ProtectfromExpropriationi = \pi_1 \cdot SettlerMortality_i + x_i'\Pi_2 + v_i \quad [1^{st} \text{ stage}]$$

$$\log(GDPpercapita_i) = \alpha \cdot ProtectfromExpropriation_i + x_i'\beta + \epsilon_i \quad [2^{nd} \text{ stage}]$$

- $ProtectfromExpropriation_i$: strength of individual property rights, proxy for the strength of institutions.
- x_i: a set of geographic controls.

High-dimensional Method

If we use high-dimensional methods with three-equation system:

$$\begin{aligned}
\log(GDPpercapita_i) &= \alpha \cdot ProtectfromExpropriation_i + x_i'\beta + \epsilon_i \\
ProtectfromExpropriationi &= \pi_1 \cdot SettlerMortality_i + x_i'\Pi_2 + v_i \\
SettlerMortality_i &= x_i'\gamma + u_i.
\end{aligned}$$

which yields three reduced form equations relating the structural variables to the controls:

$$\begin{aligned}
\log(GDPpercapita_i) &= x_i'\tilde{\beta} + \tilde{\epsilon}_i \\
ProtectfromExpropriationi &= x_i'\tilde{\Pi}_2 + \tilde{v}_i \\
SettlerMortality_i &= x_i'\gamma + u_i.
\end{aligned}$$

Steps to get valid estimation and inference for α are as follows:

1. Select controls for each of these three reduced form equations.
2. Use conventional IV estimation ($SettlerMortality_i$ as an IV for $ProtectfromExpropriation_i$), with the union of variables selected from each reduced form as included control variables.

Results

Figure: Effect of Institutions on Output (Bellonie et al. (2014), 47.)

	Latitude	*All controls*	*Double selection*
First stage	−0.5372	−0.2182	−0.5429
	(0.1545)	(0.2011)	(0.1719)
Second stage	0.9692	0.9891	0.7710
	(0.2128)	(0.8005)	(0.1971)

- 1st column only uses latitude as control in 2SLS, while 2nd column uses all controls in 2SLS.
- 3rd column is from high-dimensional method.
- The results of 1st and 3rd columns are qualitatively similar.
- From 2nd column, including all controls results in a very imprecisely estimated first-stage. The estimate of the effect of institutions is thus unreliable.

Conclusion

- To deal with high-dimensional data, dimension reduction and regularization are two ways to avoid overfitting and produce useful out-of-sample forecasts.
- LASSO performs both variable selection and regularization, which is useful for making forcasts.
- LASSO esimators tend to have substantial bias towards zero. To alleviate this, the post-LASSO is useful.
- The double-selection approach identifies potentially important confounding variables, which is useful to draw causal inference.

- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." Journal of Economic Perspectives, 28 (2): 29-50.
- Leeb, Hannes, and Benedikt M. Pötscher. 2008a. "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" Econometric Theory 24(2): 338-376.
- Leeb, Hannes, and Benedikt M. Pötscher. 2008b. "Recent Developments in Model Selection and Related Areas." Econometric Theory 24(2): 319-322.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer: New York, NY.