Mostly Dangerous Econometrics: How to do Model Selection with Inference in Mind

Victor Chernozhukov

June 25, 2015, MIT
Introduction

- Richer data and methodological developments lead us to consider more elaborate econometric models than before.
Introduction

- Richer data and methodological developments lead us to consider more elaborate econometric models than before.
- Focus discussion on the linear endogenous model

\[ y_i = d_i \alpha + \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad \mathbb{E}[\epsilon_i | x_i, z_i] = 0. \]
Introduction

- Richer data and methodological developments lead us to consider more elaborate econometric models than before.
- Focus discussion on the linear endogenous model

\[ y_i = d_i \alpha + \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i, \]  

\[ \mathbb{E}[\epsilon_i | x_i, z_i] = 0. \]

- Controls can be richer as more features become available (Census characteristics, housing characteristics, geography, text data)

\[ \Leftarrow \text{“big” data} \]
Introduction

- Richer data and methodological developments lead us to consider more elaborate econometric models than before.
- Focus discussion on the linear endogenous model

\[
y_i = d_i \alpha + \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i, \tag{1}
\]

- Controls can be richer as more features become available (Census characteristics, housing characteristics, geography, text data)
  \(\leftarrow \) “big” data
- Controls can contain transformation of “raw” controls in an effort to make models more flexible
  \(\leftarrow \) nonparametric series modeling, “machine learning”
Introduction

- This **forces** us to explicitly consider **model selection** to select controls that are “most relevant”.
- Model selection techniques:
  - **CLASSICAL**: *t* and *F* tests
  - **MODERN**: *Lasso*, Regression Trees, Random Forests, Boosting
Introduction

- This **forces** us to explicitly consider **model selection** to select controls that are “most relevant”.

- Model selection techniques:
  - **CLASSICAL**: t and F tests
  - **MODERN**: Lasso, Regression Trees, Random Forests, Boosting

If you are using *any* of these MS techniques directly in (1), you are doing it *wrong*.

Have to do *additional selection* to make it right.
An Example: Effect of Institutions on the Wealth of Nations

- Acemoglu, Johnson, Robinson (2001)
- Impact of institutions on wealth

\[ y_i = d_i + \alpha + \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i, \]  

Instrument \( z_i \): the early settler mortality (200 years ago)

Sample size \( n = 67 \)

Specification of controls:
- Basic: constant, latitude (\( p=2 \))
- Flexible: + cubic spline in latitude, continent dummies (\( p=16 \))
### Example: The Effect of Institutions

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Effect</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Controls</td>
<td>.96**</td>
<td>0.21</td>
</tr>
<tr>
<td>Flexible Controls</td>
<td>.98</td>
<td>0.80</td>
</tr>
</tbody>
</table>

▶ Is it ok to drop the additional controls?
## Example: The Effect of Institutions

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Effect</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Controls</td>
<td>0.96**</td>
<td>0.21</td>
</tr>
<tr>
<td>Flexible Controls</td>
<td>0.98</td>
<td>0.80</td>
</tr>
</tbody>
</table>

- Is it ok to drop the additional controls?

  Potentially Dangerous.
**Example: The Effect of Institutions**

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Effect</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Controls</td>
<td>0.96**</td>
<td>0.21</td>
</tr>
<tr>
<td>Flexible Controls</td>
<td>0.98</td>
<td>0.80</td>
</tr>
</tbody>
</table>

- Is it ok to drop the additional controls?

Potentially Dangerous. Very.
Analysis: things can go wrong even with \( p = 1 \)

- Consider a very simple exogenous model

\[
y_i = d_i \alpha + x_i \beta + \epsilon_i, \quad \mathbb{E}[\epsilon_i | d_i, x_i] = 0.
\]

- Common practice is to do the following.

**Post-single selection** procedure:

**Step 1.** Include \( x_i \) only if it is a significant predictor of \( y_i \) as judged by a conservative test (t-test, Lasso, etc.). Drop it otherwise.

**Step 2.** Refit the model after selection, use standard confidence intervals.

- This can **fail miserably**, if \( |\beta| \) is close to zero but not equal to zero, formally if

\[
|\beta| \propto \frac{1}{\sqrt{n}}
\]
What can go wrong? Distribution of $\sqrt{n}(\hat{\alpha} - \alpha)$ is not what you think

$$y_i = d_i \alpha + x_i \beta + \epsilon_i, \quad d_i = x_i \gamma + v_i$$

$\alpha = 0, \quad \beta = .2, \quad \gamma = .8,$

$n = 100$

$\epsilon_i \sim N(0, 1)$

$(d_i, x_i) \sim N \left(0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix} \right)$

- selection done by a t-test

Reject $H_0 : \alpha = 0$ (the truth) about 50% of the time (with nominal size of 5%)
What can go wrong? Distribution of $\sqrt{n}(\hat{\alpha} - \alpha)$ is not what you think

$y_i = d_i \alpha + x_i \beta + \epsilon_i, \quad d_i = x_i \gamma + v_i$

$\alpha = 0, \quad \beta = .2, \quad \gamma = .8,$

$n = 100$

$\epsilon_i \sim N(0, 1)$

$(d_i, x_i) \sim N \left(0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix} \right)$

selection done by Lasso

Reject $H_0 : \alpha = 0$ (the truth) of no effect about 50% of the time
Solutions?

Pseudo-solutions:

- **Practical**: bootstrap (does not work),
- **Classical**: assume the problem away by assuming that either $\beta = 0$ or $|\beta| \gg 0$,
- **Conservative**: don’t do selection
Solution: Post-double selection

- **Post-double selection** procedure (BCH, 2010, ES World Congress, ReStud, 2013):

  **Step 1.** Include $x_i$ if it is a significant predictor of $y_i$ as judged by a conservative test (t-test, Lasso etc).

  **Step 2.** Include $x_i$ if it is a significant predictor of $d_i$ as judged by a conservative test (t-test, Lasso etc). [In the IV models must include $x_i$ if it a significant predictor of $z_i$].

  **Step 3.** Refit the model after selection, use standard confidence intervals.

**Theorem (Belloni, Chernozhukov, Hansen: WC ES 2010, ReStud 2013)**

*DS works in low-dimensional setting and in high-dimensional approximately sparse settings.*
Double Selection Works

\[ y_i = d_i \alpha + x_i \beta + \epsilon_i, \quad d_i = x_i \gamma + v_i \]
\[ \alpha = 0, \quad \beta = .2, \quad \gamma = .8, \]
\[ n = 100 \]
\[ \epsilon_i \sim N(0, 1) \]
\[ (d_i, x_i) \sim N \left(0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix} \right) \]

► double selection
  done by t-tests

Reject \( H_0 : \alpha = 0 \) (the truth) about 5% of the time (for nominal size = 5%)
Double Selection Works

\[ y_i = d_i \alpha + x_i \beta + \epsilon_i, \quad d_i = x_i \gamma + v_i \]

\[ \alpha = 0, \quad \beta = .2, \quad \gamma = .8, \]

\[ n = 100 \]

\[ \epsilon_i \sim N(0, 1) \]

\[ (d_i, x_i) \sim N \left( 0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix} \right) \]

\[ \text{Reject } H_0 : \alpha = 0 \text{ (the truth) about 5\% of the time (nominal size = 5\%)} \]
Intuition

- The **Double Selection** — the selection among the controls $x_i$ that predict *either* $d_i$ *or* $y_i$ — 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 model selection version of Frisch-Waugh-Lovell partialling-put procedure for estimating linear regression.

- The double selection method is robust to moderate selection mistakes in the two selection steps.
More Intuition via OMVB Analysis

Think about omitted variables bias:

\[ y_i = \alpha d_i + \beta x_i + \zeta_i \; ; \; d_i = \gamma x_i + v_i \]

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

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

- the good term is asymptotically normal, and we want \( \sqrt{n}\gamma\beta \rightarrow 0 \).

- **single selection** can drop \( x_i \) only if \( \beta = O(\sqrt{1/n}) \), but
  \[ \sqrt{n}\gamma\sqrt{1/n} \not\rightarrow 0 \]

- **double selection** can drop \( x_i \) only if both \( \beta = O(\sqrt{1/n}) \) and \( \gamma = O(\sqrt{1/n}) \), that is, if
  \[ \sqrt{n}\gamma\beta = O(1/\sqrt{n}) \rightarrow 0. \]
Example: The Effect of Institutions, Continued

Going back to Acemoglu, Johnson, Robinson (2001):

- **Double Selection**: include $x_{ij}$’s that are significant predictors of either $y_i$ or $d_i$ or $z_i$, as judged by Lasso. Drop otherwise.

<table>
<thead>
<tr>
<th>Intitutions</th>
<th>Effect</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Controls</td>
<td>.96**</td>
<td>0.21</td>
</tr>
<tr>
<td>Flexible Controls</td>
<td>.98</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>Double Selection</strong></td>
<td>.78**</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Application: Effect of Abortion on Murder Rates in the U.S.

Estimate the consequences of abortion rates on crime in the U.S., Donohue and Levitt (2001)

\[ y_{it} = \alpha d_{it} + x'_{it}\beta + \zeta_{it} \]

- \( y_{it} \) = change in crime-rate in state \( i \) between \( t \) and \( t - 1 \),
- \( d_{it} \) = change in the (lagged) abortion rate,
1. \( x_{it} \) = basic controls (time-varying confounding state-level factors, trends; \( p = 20 \))
2. \( x_{it} \) = flexible controls (basic + state initial conditions + two-way interactions of all these variables)
- \( p = 251, n = 576 \)
## Effect of Abortion on Murder, continued

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Abortion on Murder Effect</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Controls</td>
<td>-0.204**</td>
<td>0.068</td>
</tr>
<tr>
<td>Flexible Controls</td>
<td>-0.321</td>
<td>1.109</td>
</tr>
<tr>
<td>Single Selection</td>
<td>-0.202**</td>
<td>0.051</td>
</tr>
<tr>
<td>Double Selection</td>
<td>-0.166</td>
<td>0.216</td>
</tr>
</tbody>
</table>

- Double selection by Lasso: 8 controls selected, including state initial conditions and trends interacted with initial conditions
This is sort of a negative result, unlike in AJR (2011)

Double selection does not always overturn results. Plenty of positive results confirming:

- Barro and Lee’s convergence results in cross-country growth rates;
- Poterba et al results on positive impact of 401(k) on savings;
- Acemoglu et al (2014) results on democracy causing growth;
High-Dimensional Prediction Problems

- **Generic prediction problem**

  \[ u_i = \sum_{j=1}^{p} x_{ij} \pi_j + \zeta_i, \quad \mathbb{E}[\zeta_i \mid x_i] = 0, \quad i = 1, \ldots, n, \]

  can have \( p = p_n \) small, \( p \propto n \), or even \( p \gg n \).

- In the double selection procedure, \( u_i \) could be outcome \( y_i \), treatment \( d_i \), or instrument \( z_i \). Need to find good predictors among \( x_{ij} \)'s.

- **Approximate Sparsity**: after sorting, absolute values of coefficients decay fast enough:

  \[ |\pi|_{(j)} \leq A j^{-a}, \quad a > 1, j = 1, \ldots, p = p_n, \forall n \]

- **Restricted Isometry**: small groups of \( x_{ij} \)'s are not close to being collinear.
Selection of Predictors by Lasso

Assuming $x_{ij}'s$ normalized to have the second empirical moment to 1.

- **Ideal (Akaike, Schwarz):** minimize
  \[
  \sum_{i=1}^{n} \left( u_i - \sum_{j=1}^{p} x_{ij} b_j \right)^2 + \lambda \left( \sum_{j=1}^{p} 1\{b_j \neq 0\} \right).
  \]

- **Lasso (Bickel, Ritov, Tsybakov, Annals, 2009):** minimize
  \[
  \sum_{i=1}^{n} \left( u_i - \sum_{j=1}^{p} x_{ij} b_j \right)^2 + \lambda \left( \sum_{j=1}^{p} |b_j| \right), \quad \lambda = \sqrt{E \zeta^2} 2^{\sqrt{2n \log(pn)}}
  \]
Selection of Predictors by Lasso

Assuming $x_{ij}'$s normalized to have the second empirical moment to 1.

- **Ideal (Akaike, Schwarz):** minimize
  \[
  \sum_{i=1}^{n} \left( u_i - \sum_{j=1}^{p} x_{ij} b_j \right)^2 + \lambda \left( \sum_{j=1}^{p} 1\{b_j \neq 0\} \right).
  \]

- **Lasso (Bickel, Ritov, Tsybakov, Annals, 2009):** minimize
  \[
  \sum_{i=1}^{n} \left( u_i - \sum_{j=1}^{p} x_{ij} b_j \right)^2 + \lambda \left( \sum_{j=1}^{p} |b_j| \right), \quad \lambda = \sqrt{\mathbb{E} \zeta^2} 2 \sqrt{2n \log(pn)}
  \]

- **Root Lasso (Belloni, Chernozhukov, Wang, Biometrika, 2011):** minimize
  \[
  \sqrt{\sum_{i=1}^{n} \left( u_i - \sum_{j=1}^{p} x_{ij} b_j \right)^2} + \lambda \left( \sum_{j=1}^{p} |b_j| \right), \quad \lambda = \sqrt{2n \log(pn)}
  \]
Lasso provides high-quality model selection

Theorem (Belloni and Chernozhukov: Bernoulli, 2013, Annals, 2014)

Under approximate sparsity and restricted isometry conditions, Lasso and Root-Lasso find parsimonious models of approximately optimal size

\[ s = n^{2a}. \]

Using these models, the OLS can approximate the regression functions at the nearly optimal rates in the root mean square error:

\[ \sqrt{\frac{s}{n}} \log(pn) \]
Double Selection in Approximately Sparse Regression

- **Exogenous model**

  \[ y_i = d_i \alpha + \sum_{j=1}^{p} x_{ij} \beta_j + \zeta_i, \quad \mathbb{E}[\zeta_i \mid d_i, x_i] = 0, \quad i = 1, \ldots, n, \]

  \[ d_i = \sum_{j=1}^{p} x_{ij} \gamma_j + \nu_i, \quad \mathbb{E}[\nu_i \mid x_i] = 0, \quad i = 1, \ldots, n, \]

  can have \( p \) small, \( p \propto n \), or even \( p \gg n \).

- **Approximate Sparsity**: after sorting absolute values of coefficients decay fast enough:

  \[ |\beta|_{(j)} \leq A j^{-a}, \quad a > 1, \quad |\gamma|_{(j)} \leq A j^{-a}, \quad a > 1. \]

- **Restricted Isometry**: small groups of \( x'_{ij}s \) are not close to being collinear.
Double Selection Procedure

▶ **Post-double selection** procedure (BCH, 2010, ES World Congress, ReStud 2013):

**Step 1.** Include $x_{ij}$’s that are significant predictors of $y_i$ as judged by LASSO or OTHER high-quality selection procedure.

**Step 2.** Include $x_{ij}$’s that are significant predictors of $d_i$ as judged by LASSO or OTHER high-quality selection procedures.

**Step 3.** Refit the model by least squares after selection, use standard confidence intervals.
Uniform Validity of the Double Selection

Theorem (Belloni, Chernozhukov, Hansen: WC 2010, ReStud 2013)

*Uniformly within a class of approximately sparse models with restricted isometry conditions*

\[ \sigma_n^{-1} \sqrt{n(\hat{\alpha} - \alpha_0)} \rightarrow_d N(0, 1), \]

where \( \sigma_n^2 \) is conventional variance formula for least squares. Under homoscedasticity, semi-parametrically efficient.

- Model selection mistakes are asymptotically negligible due to double selection.
- Analogous result also holds for endogenous models, see Chernozhukov, Hansen, Spindler, *Annual Review of Economics*, 2015.
Monte Carlo Confirmation

- In this simulation we used: $p = 200, \ n = 100, \ \alpha_0 = .5$

\[ y_i = d_i \alpha + x_i' \beta + \zeta_i, \ \zeta_i \sim N(0,1) \]

\[ d_i = x_i' \gamma + v_i, \ v_i \sim N(0,1) \]

- approximately sparse model:

\[ |\beta_j| \propto 1/j^2, \ |\gamma_j| \propto 1/j^2 \]

- $R^2 = .5$ in each equation

- regressors are correlated Gaussians:

\[ x \sim N(0, \Sigma), \ \Sigma_{kj} = (0.5)^{|j-k|}. \]
Distribution of Post Double Selection Estimator

\[ p = 200, \ n = 100 \]
Distribution of Post-Single Selection Estimator

\[ p = 200 \text{ and } n = 100 \]
Generalization: Orthogonalized or “Doubly Robust” Moment Equations

- **Goal:**
  - inference on structural parameter $\alpha$ (e.g., elasticity)
  - having done Lasso & friends fitting of reduced forms $\eta(\cdot)$

- Use orthogonalization methods to remove biases. This often amounts to solving auxiliary prediction problems.

- In a nutshell, we want to set up moment conditions

  $$
  \mathbb{E}[g(W_{\text{data}}, \alpha_0, \eta_0)] = 0
  $$

  such that the orthogonality conditions hold:

  $$
  \partial_{\eta} \mathbb{E}[g(W, \alpha_0, \eta)] \bigg|_{\eta=\eta_0} = 0
  $$

- See my website for papers on this.
Inference on Structural/Treatment Parameters

Without Orthogonalization

With Orthogonalization
Conclusion

- It is time to address model selection
- Mostly dangerous: naive (post-single) selection does not work
- Double selection works
- More generally, the key is to use orthogonalized moment conditions for inference