ECON 626: Applied Microeconomics

Lecture 6:

Selection on Observables

Professors: Pamela Jakiela and Owen Ozier

Experimental and Quasi-Experimental Approaches

Approaches to causal inference (that we've discussed so far):

- The experimental ideal (i.e. RCTs)
- Natural experiments
- Difference-in-differences
- Instrumental variables
- Regression discontinuity

These approaches* reply on good-as-random variation in treatment; identify impact on compliers irrespective of the nature of confounds

* With possible exception of diff-in-diff

Causal Inference When All Else Fails

What can we do when we don't have an experiment or quasi-experiment?

- Credibility revolution in economics nudges us to focus on questions that can be answered through "credible" identification strategies
- Is this good for science? Is it good for humanity?

We should not restrict our attention to questions that can be answered through randomized trials, natural experiments, or quasi-experiments!

- Research frontier: using best methods available, cond'l on question

Causal Inference When All Else Fails

Non-experimental causal inference: explicit consideration of confounds

- Structural models (take a class from Sergio or Sebastian!)
- Matching estimators (just don't use propensity scores)
- Directed acyclic graphs (DAGs)
- **Coefficient stability**
- **Machine learning to select covariates**

Coefficient Stability

Motivating Example

Example: the impact of Catholic schools on high school graduation

	All Students		Catholic Elementary	
	No Controls	**w/ Controls**	**No Controls**	**w/ Controls**
Probit coefficient	0.97	0.41	0.99	1.27
S.E.	(0.17)	(0.21)	(0.24)	(0.29)
Marginal effects	[0.123]	[0.052]	[0.11]	[0.088]
Pseudo R^2	0.01	0.34	0.11	0.58

Source: Table 3 in Altonji, Elder, Taber (2005)

A Framework for Thinking About Selection Bias

$$
\begin{aligned}
Y^* &= \alpha CH + \boldsymbol{W}'\boldsymbol{\Gamma} \\
&= \alpha CH + \boldsymbol{X}'\boldsymbol{\Gamma}_X + \xi \\
&= \alpha CH + \boldsymbol{X}'\boldsymbol{\gamma} + \epsilon
\end{aligned}
$$

where

- α is the causal impact of Catholic high school (CH)
- $\boldsymbol{W}$ is all covariates, and $\boldsymbol{X}$ is observed covariates
- ϵ is defined to be orthogonal to $\boldsymbol{X}$ s.t. $Cov(X, \epsilon) = 0$

In this framework, why is the OLS estimate of α biased?

How Severe Is Selection on Unobservables?

Consider a linear projection of CH onto $\boldsymbol{X}'\boldsymbol{\gamma}$

$$CH = \phi_0 + \phi_{X'\gamma}\boldsymbol{X}'\boldsymbol{\gamma} + \phi_\epsilon \epsilon$$

Typical identification assumption in OLS: $\phi_\epsilon = 0$

- AET propose weaker proportional selection condition: $\phi_\epsilon = \phi_{X'\gamma}$

Proportional selection is equivalent to following condition:

$$\frac{E[\epsilon|CH=1] - E[\epsilon|CH=0]}{Var(\epsilon)} = \frac{E[\boldsymbol{X}'\gamma|CH=1] - E[\boldsymbol{X}'\gamma|CH=0]}{Var(\boldsymbol{X}'\gamma)}$$

Let's Assume. . .

1. Elements of $\boldsymbol{X}$ chosen at random W that determine $\boldsymbol{W}$
2. $\boldsymbol{X}$ and $\boldsymbol{W}$ have many elements; none dominant predictors of Y
3. Additional (apparently hard to state) assumption:

   "Roughly speaking, the assumption is that the regression of CH^* on Y^* - αCH is equal to the regression of the part of CH^* that is orthogonal to $\boldsymbol{X}$ on the corresponding part of Y^* - αCH."

   where CH^* is an unobserved latent variable that determines CH

Bounding Selection on Unobservables

Define $CH = \boldsymbol{X}'\boldsymbol{\beta} + \widetilde{CH}$ and re-write estimating equation:

$$Y^* = \alpha \widetilde{CH} + \boldsymbol{X}'(\boldsymbol{\gamma} + \alpha\boldsymbol{\beta}) + \epsilon$$

This gives us a formula for selection bias:

$$plim\ \hat{\alpha} = \alpha + \frac{Var(CH)}{Var(\widetilde{CH})} \left(E[\epsilon|CH = 1] - E[\epsilon|CH = 0] \right)$$

The bias is bounded under proportional selection assumption:

$$E[\epsilon|CH = 1] - E[\epsilon|CH = 0] = Var(\epsilon) \cdot \frac{E[\boldsymbol{X}'\gamma|CH = 1] - E[\boldsymbol{X}'\gamma|CH = 0]}{Var(\boldsymbol{X}'\gamma)}$$

Some Restrictions Apply

"Note that when $Var(\epsilon)$ is very large relative to $Var(\boldsymbol{X}'\gamma)$,
what one can learn is limited

...

even a small shift in $\left(E[\epsilon|CH=1]-E[\epsilon|CH=0]\right)/Var(\epsilon)$
is consistent with a large bias in α."

The degree of selection bias is bounded, but bounds may be wide:

$$\text{bias} < \frac{Var(CH)}{Var(\widetilde{CH})}\left(Var(\epsilon)\cdot\frac{E[\boldsymbol{X}'\gamma|CH=1]-E[\boldsymbol{X}'\gamma|CH=0]}{Var(\boldsymbol{X}'\gamma)}\right)$$

Altonji, Elder, Taber (2005)

TABLE 3
OLS and Probit Estimates of Catholic High School Effects in Subsamples of NELS:88 (Weighted)

	Full Sample: Controls				Catholic 8th Grade Attendees: Controls			
	None (1)	Family Background, City Size, and Region[a] (2)	Col. 2 Plus 8th Grade Tests (3)	Col. 3 Plus Other 8th Grade Measures[b] (4)	None (5)	Family Background, City Size, and Region[a] (6)	Col. 2 Plus 8th Grade Tests (7)	Col. 3 Plus Other 8th Grade Measures[b] (8)
A. High School Graduation								
Probit	.97	.57	.48	.41	.99	.88	.95	1.27
	(.17)	(.19)	(.22)	(.21)	(.24)	(.25)	(.27)	(.29)
	[.123]	[.081]	[.068]	[.052]	[.105]	[.084]	[.081]	[.088]
Pseudo R^2[c]	.01	.16	.21	.34	.11	.35	.44	.58
B. College in 1994								
Probit	.73	.37	.33	.32	.60	.48	.56	.60
	(.08)	(.09)	(.09)	(.09)	(.13)	(.15)	(.15)	(.15)
	[.283]	[.106]	[.084]	[.074]	[.236]	[.154]	[.154]	[.149]
Pseudo R^2	.02	.19	.29	.34	.04	.18	.29	.36
C. 12th Grade Reading Score								
OLS	4.28	2.08	1.18	1.14	1.92	.17	.37	.33
	(.47)	(.54)	(.38)	(.38)	(.82)	(.98)	(.63)	(.62)
R^2	.01	.19	.60	.60	.01	.19	.59	.62
D. 12th Grade Math Score								
OLS	4.86	1.98	1.07	.92	2.79	1.10	1.46	1.14
	(.44)	(.54)	(.34)	(.32)	(.77)	(1.00)	(.53)	(.46)
R^2	.01	.26	.72	.74	.02	.26	.73	.77

Altonji, Elder, Taber (2005)

TABLE 3
OLS and Probit Estimates of Catholic High School Effects in Subsamples of NELS:88 (Weighted)

		Catholic 8th Grade Attendees: Controls			
		None (5)	Family Background, City Size, and Region[a] (6)	Col. 2 Plus 8th Grade Tests (7)	Col. 3 Plus Other 8th Grade Measures[b] (8)
A. High School Graduation	Probit	.99	.88	.95	1.27
		(.24)	(.25)	(.27)	(.29)
		[.105]	[.084]	[.081]	[.088]
	Pseudo R^{2c}	.11	.35	.44	.58
B. College in 1994	Probit	.60	.48	.56	.60
		(.13)	(.15)	(.15)	(.15)
		[.236]	[.154]	[.154]	[.149]
	Pseudo R^2	.04	.18	.29	.36
C. 12th Grade Reading Score	OLS	1.92	.17	.37	.33
		(.82)	(.98)	(.63)	(.62)
	R^2	.01	.19	.59	.62
D. 12th Grade Math Score	OLS	2.79	1.10	1.46	1.14
		(.77)	(1.00)	(.53)	(.46)
	R^2	.02	.26	.73	.77

Altonji, Elder, Taber (2005)

TABLE 6

AMOUNT OF SELECTION ON UNOBSERVABLES RELATIVE TO SELECTION ON OBSERVABLES REQUIRED TO ATTRIBUTE THE ENTIRE CATHOLIC SCHOOL EFFECT TO SELECTION BIAS

Outcome	$[\hat{E}(\boldsymbol{X}'\hat{\boldsymbol{\gamma}} \mid CH = 1) - \hat{E}(\boldsymbol{X}'\hat{\boldsymbol{\gamma}} \mid CH = 0)] \div \widehat{\mathrm{Var}}(\boldsymbol{X}'\hat{\boldsymbol{\gamma}})$ (1)	$\widehat{\mathrm{Var}}(\hat{\epsilon})$ (2)	$E(\epsilon \mid CH = 1) - E(\epsilon \mid CH = 0)$[a] (3)	$\mathrm{Cov}(\epsilon, \widetilde{CH}) \div \mathrm{Var}(\widetilde{CH})$ (4)	$\hat{\alpha}$ (5)	Implied Ratio[b] (6)
	A. $\hat{\alpha}$ Estimated from the Catholic Eighth Grade Subsample, Full Set of Controls[c]					
High school graduation (N=859)	.24	1.00	.24	.29	1.03 (.31)	3.55
College attendance (N=834)	.39	1.00	.39	.47	.67 (.16)	1.43
12th grade reading (N=739)	.091	36.00	3.28	3.94	.33 (.62)	.08
12th grade math (N=739)	.038	24.01	.91	1.09	1.14 (.46)	1.04

Bellows and Miguel (2009)

Table 3
Community meetings and conflict victimization.

	Dependent variable: did you attend any community meetings in the past year?			
	IRCBP			
	2005 and 2007		2007	
Explanatory variables	(1)	(2)	(3)	(4)
Conflict victimization index	0.0704***	0.0652***	0.0775***	0.0686***
	(0.0164)	(0.0165)	(0.0253)	(0.0246)
Respondent is female		−0.1300***		−0.1276***
		(0.0084)		(0.0126)
Respondent age		0.0003		0.0002
		(0.0003)		(0.0005)
Respondent has any education		0.0590***		0.0466**
		(0.0108)		(0.0194)
Traditional authority household		0.0928***		0.0647***
		(0.0128)		(0.0194)
1990 Household head had any education				0.0205
				(0.0199)
1990 Household had a traditional leader				0.1054***
				(0.0217)
1990 Household had a community leader				−0.0067
				(0.0169)
R-squared	0.361	0.391	0.267	0.298
Observations	10,471	10,471	5193	5193
Enumeration area/Year fixed effects	X	X	X	X

Oster (2019): A Practical Applications of AET

"A common approach to evaluating robustness to omitted variable bias is to observe coefficient movements after inclusion of controls. This is informative only if selection on observables is informative about selection on unobservables. Although this link is known in theory (i.e. Altonji, Elder and Taber 2005), very few empirical papers approach this formally. I develop an extension of the theory which connects bias explicitly to coefficient stability. I show that it is necessary to take into account coefficient and R-squared movements. I develop a formal bounding argument. I show two validation exercises and discuss application to the economics literature."

Oster (2019): A Practical Applications of AET

Given a treatment T, define the proportional selection coefficient:

$$\delta = \frac{Cov(\epsilon, T)}{Var(\epsilon)} \Big/ \frac{Cov(\boldsymbol{X}'\gamma, T)}{Var(\boldsymbol{X}'\gamma)}$$

Then:

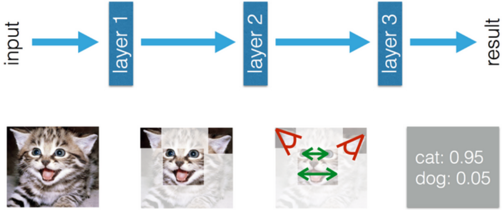
$$\beta^* \approx \tilde{\beta} - \delta \left[\overset{o}{\beta} - \tilde{\beta} \right] \frac{R_{max} - \tilde{R}}{\tilde{R} - \overset{o}{R}} \xrightarrow{p} \beta$$

where:

- $\overset{o}{\beta}$ and $\overset{o}{R}$ are from a univariate regression of Y on T
- $\tilde{\beta}$ and $\tilde{R}$ are from a regression including controls
- R_{max} is the maximum achievable R^2 (possible 1)

Very Simple Machine Learning

What Is Machine Learning?



What Is Machine Learning?

A set of extensions to the standard econometric toolkit (read: "OLS") aimed at improving predictive accuracy, particularly w/ many variables

- Subset selection
- Shrinkage (LASSO, Ridge regression)
- Regression trees, random forests

Machine learning introduces new tools, relabels existing tools

- **training data/sample/examples**: your data
- **features**: independent variables, covariates

Main focus is on predicting Y, not testing hypotheses about β

$\Rightarrow$ ML "results" about β may not be robust

Can We Improve on OLS?

A standard linear model is not (always) the best way to predict Y:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Can we improve on OLS?

- When $p > N$, OLS is not feasible
- When p is large relative to N, model may be prone to over-fitting
- OLS explains both structural and spurious relationships in data

Extensions to OLS identify "strongest" predictors of Y

- Strength of correlation vs. (out-of-sample) robustness

Assumption: exact or approximate **sparcity**

Best Subset Selection

A **best subset selection** algorithm:

- For each $k = 1, 2, \ldots, p$
  - Fit all models containing exactly k covariates
  - Identify the "best" in terms of R^2
- Choose the **best subset** based on cross-validation, adjusted R^2, etc.
  - Need to address the fact that R^2 always increases with k

When p is large, best subset selection is not feasible

Alternatives to Best Subset Selection

A **backward stepwise selection** algorithm:

- Start with the "full" model containing p covariates
- At each step, drop one variable
  - Choose the variable the minimizes decline in R^2
- Choose among "best" subsets of covariates thus identified (conditional on $k \leq p$) using cross-validation, adjusted R^2, etc.

Alternatives to Best Subset Selection

An even simpler **backward stepwise selection** algorithm:

- Start with the full model containing p covariates
- Drop covariates with p-values below 0.05
- Re-estimate, repeat until all covariates are statistically significant

Stepwise selection algorithm's may or may not yield optimal covariates

- When variables are not independent/orthogonal, how much one variable matters can depend on which other variables are included

Best Subset Selection

In OLS, we seek to minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

Best subset selection can be expressed as: choose β to minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p} I\left(\beta_j \neq 0\right) \leq s$$

where s is the number of regressors/predictors/features/covariates

$\Rightarrow$ But we solve it algorithmically, not analytically

$\Rightarrow$ When p is large, finding the best subset is hard

LASSO and Ridge Regression

Ridge regression solves a closely related minimization problem:

$$\min_\beta \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p}\beta_j^2 \leq s$$

or, equivalently,

$$\min_\beta \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for some **tuning parameter** $\lambda \geq 0$

Ridge regression shrinks OLS coefficients toward zero

- Shrinkage is more or less proportional, so ridge regression does not identify a subset of regressors to include/retain in analysis/prediction

LASSO and Ridge Regression

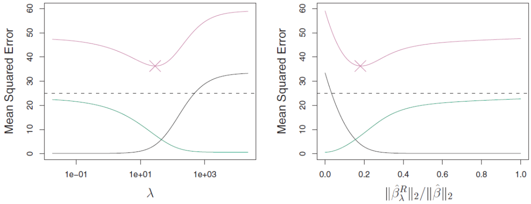


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Gauss-Markov Theorem: OLS is best linear unbiased estimator (BLUE)

- Estimators that are (a little) biased can generate better predictions

LASSO and Ridge Regression

LASSO (Least Absolute Shrinkage and Selection Operator):

$$\min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

for some **tuning parameter** $\lambda \geq 0$

LASSO combines benefits of subset selection, ridge regression

- Less computationally intensive than subset selection
- Sets some coefficients to 0 $\rightarrow$ identifies parsimonious model
- Better than ridge regression when most covariates are garbage

LASSO and Ridge Regression

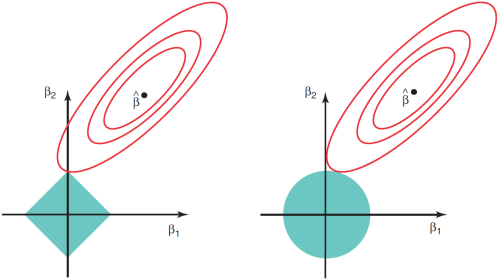


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$*, while the red ellipses are the contours of the RSS.*

LASSO constraint region has sharp corners $\Rightarrow$ some coefficients set to 0

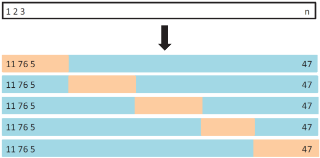
Three Approaches to Choosing λ (1/3)

Statistics based on in-sample fit:

- Function of n, RSS, plus degrees of freedom correction
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
  - Extended Bayesian Information Criterion (EBIC)
- Default implemented by Stata's `lasso2` command

These approaches tend to choose "too many" variables when n is small

Three Approaches to Choosing λ (2/3)



k-fold cross-validation

- Randomly sort observations in k groups
- For each group k, estimate LASSO on on rest of sample and predict MSE using observations in k; average to get MSE(λ)
- Iterate over λ values to choose optimal λ

Three Approaches to Choosing λ (3/3)

Belloni et al. (2012): alternative approach to choosing λ

- Relies on assumption of approximate sparsity
- Chooses λ iteratively based on data
- Allows for heteroskedasticity

Three approaches may generate very different sets of controls

- AIC may allow for too many controls when p is large
- Rigorous methods may suggest no controls are needed!
- Costs of too many/too few may vary across empirical contexts

Using Stata's `lasso2` Command

```
. lasso2 Y A1 A2 A3 A4 B1 B2 B3 B4 C1 C2 C3 C4

  Knot |  ID      Lambda      s       L1-Norm        EBIC      R-sq   | Entered/removed
-------+-----------------------------------------------------------------+----------------
     1 |   1   274.69944      1       0.00000     373.99443   0.0000   | Added _cons.
     2 |   2   250.29590      2       0.05750     376.74198   0.0127   | Added A3.
     3 |   3   228.06030      3       0.12916     379.14887   0.0268   | Added B1.
     4 |   5   189.33967      4       0.34741     376.64324   0.0641   | Added B2.
     5 |   7   157.19312      5       0.59797     374.32176   0.0991   | Added A4.
     6 |   9   130.50449      6       0.88119     372.19232   0.1319   | Added A1.
     7 |  14    81.96062      7       1.47908     365.26203   0.1834   | Added C2.
     8 |  15    74.67947      8       1.59405     368.76792   0.1907   | Added C3.
     9 |  16    68.04515      9       1.72958     372.12859   0.1985   | Added A2.
    10 |  17    62.00020     10       1.86039     375.70853   0.2054   | Added B4.
    11 |  18    56.49228     11       1.99359     379.40878   0.2117   | Added C1.
    12 |  22    38.93794     12       2.45073     380.21448   0.2292   | Added C4.
    13 |  50     2.87779     13       3.47700     380.98320   0.2464   | Added B3.
Use 'long' option for full output. Type e.g. 'lasso2, lic(ebic)' to run the model selected
> by EBIC.
```

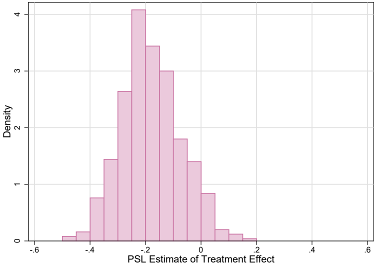
Using Stata's `lasso2` Command

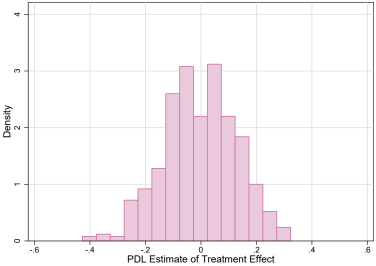
```
. lasso2, lic(ebic)

Use lambda=89.9516808401893 (selected by EBIC).
```

Selected	Lasso	Post-est OLS
A1	**0.1420802**	**0.3854609**
A3	**0.4510036**	**0.6742726**
A4	**0.1905700**	**0.4034061**
B1	**0.3653137**	**0.6000835**
B2	**0.2291416**	**0.4171147**
Partialled-out*		
_cons	**0.1234630**	**0.0877426**

Post-Double-LASSO Estimation





Using LASSO to address selection bias through **post-double-selection**:

- Using LASSO to select covariates that predict/explain Y leads to biased estimates of treatment effects of T (Belloni et al. 2014)
- PDL: use LASSO to predict Y and T, include **all** chosen controls