# Lecture 18: Lasso for Causal Inference

## Big Data and Machine Learning for Applied Economics
## Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

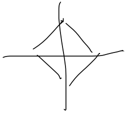
October 14, 2021

# Agenda

# Recap: Regularization

- For $\lambda \geq 0$ given, consider minimizing the following objective function
- Lasso:

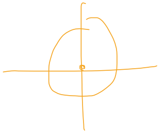
$$min_\beta L(\beta) = \underbrace{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2}_{} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{} \quad (1)$$



- Ridge:

$\cancel{+\lambda \sum (\beta_j)^p}$

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



$\sum_{j=1}^{p} (\beta_j)^2 = \sum_{j=1}^{p} (\beta_j - 0)^2$

$\lambda \sum (\beta_j - 0)^2 = \sum (\sqrt{\lambda}\beta_j - 0)^2 = \sum (x_j \beta_j - y)^2 = \sum (y - x_j\beta_j)^2$

$x_j = (0, \ \sqrt{\lambda}, \ -0)$

$y_j = 0$

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

# Recap: Regularization

$K > n$

- Elastic net: happy medium.
  - Good job at prediction and selecting variables

$n\ P$

$$= \sum (y_i - \beta_0 - \sum x_{ij}\beta_j)^2 + \lambda_1 \sum |\beta_j|$$

$$min_{\beta} NEL(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \underbrace{\lambda \sum_{j=1}^{p}(\beta_j)^2}_{\text{Ridge}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{lasso}} \tag{3}$$

- Mixes Ridge and Lasso
- Lasso selects predictors
- Strict convexity part of the penalty (ridge) solves the grouping instability problem
- H.W.: $\beta_{OLS} > 0$ one predictor standarized

$\beta < 0$

$$\hat{\beta}_{naive\,EN} = \frac{\overbrace{\left(\hat{\beta}_{OLS} - \frac{\lambda_1}{2}\right)_+}^{\text{lasso}}}{1 + \lambda_2} \tag{4}$$

# Elastic Net

- Elastic Net: reescaled version of Naive version
- Double Shrinkage introduces "too" much bias, *final* version "corrects" for this

$$\hat{\beta}_{EN} = \frac{1}{\sqrt{1+\lambda_2}} \hat{\beta}_{naive\,EN} \tag{5}$$

- Careful sometimes software asks.
- How to choose $(\lambda_1, \lambda_2)$? $\rightarrow$ Bidimensional Crossvalidation
- Zou, H. & Hastie, T. (2005)

# Model Selection When the Goal is Causal Inference

## Motivation

- In this course our objective is prediction

# Model Selection When the Goal is Causal Inference

## Motivation

- In this course our objective is prediction
- But since we are economists, inference is always there

# Model Selection When the Goal is Causal Inference

## Motivation

- In this course our objective is prediction
- But since we are economists, inference is always there
- Can we use some of these models to do causal inference?
- We are going to see how we can use lasso when inference is the main goal

# Model Selection When the Goal is Causal Inference

Let's start with the following model

outcome

$$y_i = \boxed{\alpha} D_i + \underline{g(X_i)} + \underline{\zeta_i} \tag{6}$$

were

- $D_i$ is the treatment/policy variable of interest,
- $X_i$ is a set controls
- $E[\zeta_i | D_i, X_i] = 0$

# Star experiment

$D < {\beta \atop \delta}$

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

| Explanatory variable | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Small class | 4.82 | 5.37 | 5.36 | 5.37 |
| | (2.19) | (1.26) | (1.21) | (1.19) |
| Regular/aide class | .12 | .29 | .53 | .31 |
| | (2.23) | (1.13) | (1.09) | (1.07) |
| White/Asian (1 = yes) | – | – | 8.35 | 8.44 |
| | | | (1.35) | (1.36) |
| Girl (1 = yes) | – | – | 4.48 | 4.39 |
| | | | (.63) | (.63) |
| Free lunch (1 = yes) | – | – | -13.15 | -13.07 |
| | | | (.77) | (.77) |
| White teacher | – | – | – | -.57 |
| | | | | (2.10) |
| Teacher experience | – | – | – | .26 |
| | | | | (.10) |
| Master's degree | – | – | – | -0.51 |
| | | | | (1.06) |
| School fixed effects | No | Yes | Yes | Yes |
| $R^2$ | .01 | .25 | .31 | .31 |

Note: Adapted from Krueger (1999), Table 5. The dependent variable is the Stanford Achievement Test percentile score. Robust standard errors that allow for correlated residuals within classes are shown in parentheses. The sample size is 5681.

# Model Selection When the Goal is Causal Inference

- Traditional approach: researcher selects $X_i$

# Model Selection When the Goal is Causal Inference

- Traditional approach: researcher selects $X_i$
- Problem: mistakes can occur.
- Same if they use an "automatic" model selection approach.
- Why?

# Model Selection When the Goal is Causal Inference

- Traditional approach: researcher selects $X_i$
- Problem: mistakes can occur.
- Same if they use an "automatic" model selection approach.
- Why?
- It can leave out potentially important variables with small coefficients but non zero coefficients out

# Model Selection When the Goal is Causal Inference

- The omission of such variables then generally contaminates estimation and inference results based on the selected set of variables. (e.g. OVB)
- The validity of this approach is delicate because it relies on perfect model selection.
- Because model selection mistakes seem inevitable in realistic settings, it is important to develop inference procedures that are robust to such mistakes.
- Solution here: Lasso

# Model Selection When the Goal is Causal Inference

- Using Lasso is useful for prediction
- However, naively using Lasso to draw inferences about model parameters can be problematic.
- Part of the difficulty is that these procedures are designed for prediction, not for inference
- Leeb and Pötscher 2008 show that methods that tend to do a good job at prediction can lead to incorrect conclusions when inference is the main objective

# Model Selection When the Goal is Causal Inference

- Leeb and Pötscher 2008 show that methods that tend to do a good job at prediction can lead to incorrect conclusions when inference is the main
- This observation suggests that more desirable inference properties may be obtained if one focuses on model selection over the predictive parts of the economic problem
  - The reduced forms and first-stages—rather than using model selection in the structural model directly.

# Approximate sparse models

p >> n

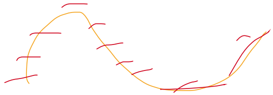
- To fix ideas suppose we have the following model and we want to predict $y$ based on $X$

$$y_i = g(X_i) + \zeta_i \tag{7}$$

with

  - $E(\zeta_i | g(x_i)) = 0$
  - $i = 1, \ldots, n$ are iid
  - To avoid over-fitting and produce good out of sample prediction we will need to restrict or regularize $g(.)$
  - Belloni's et. all approach focuses on an approach that treats $g(X_i)$ as a high-dimensional but that we can approximate linearly

# Approximate sparse models



$$g(X_i) = \sum_{j=1}^{p} \beta_j x_{ij} + r_{pi} \tag{8}$$

- where $p >> n$ and $r_{pi}$ is small enough
- Approximate sparsity of this high-dimensional linear model imposes the restriction that linear combinations of only $s << n$ $x_{ij}$ variables provide a good approximation to $g(X_i)$
- A bonus is that the identity of this $s$ $x_{ij}$ variables are a priori unknown
- And that we can have a nonzero approximation error $r_{pi}$
- We are going to try to learn the identities of these variables while estimating the coefficients.

# Approximate sparse models

- We can use Lasso that is slightly modified

$$L(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=2}^{p} |\beta_j| \boxed{\gamma_j} \tag{9}$$

  - where $\lambda > 0$ is the penalty level chosen using Belloni, Chen, Chernozhukov, and Hansen (2012)

    $\lambda = 2c\sqrt{n}\,\sigma\,\Phi^{-1}\left(1-\frac{\gamma}{2p}\right)$
  - $\gamma_j$ are *penalty loadings* $E[x^2]$
  - *penalty loadings* are chosen to insure equivariance of coefficient estimates to rescaling of $x_{ij}$ and can also be chosen to address heteroskedasticity, clustering, and non-gaussian errors

    $E(x^2\varepsilon^2)$

# Inference with Selection among Many Controls

$g(x_i) = x_i \theta_y + r_{yi}$

- Under the approximate sparse models assumption
- We consider a linear model where a treatment variable, $D_i$, is taken as exogenous after conditioning on control variables

$$y_i = \alpha D_i + \overbrace{X_i'\theta_y + r_{yi}}^{g(x_i)} + \zeta_i \tag{10}$$

- where $E[\zeta_i | d_i, x_i, r_{yi}] = 0$
- $X_i$ is a p-dimensional vector with $p >> n$, but approximately sparse
- $r_{yi}$ is an approximation error
- the parameter of interest is $\alpha$

# Inference with Selection among Many Controls

- Naive approach

$$y_i = \alpha D_i + X_i'\theta_y + r_{yi} + \zeta_i \tag{11}$$

  - Select control variables by applying Lasso, forcing the treatment variable to remain in the model
  - One could then try to estimate and do inference about $\alpha$ by applying ordinary least squares with $y_i$ as the outcome, and $D_i$ and any selected control variables as regressors.

# Inference with Selection among Many Controls

- Naive approach

$$y_i = \alpha D_i + X_i'\theta_y + r_{yi} + \zeta_i \tag{11}$$

  - Select control variables by applying Lasso, forcing the treatment variable to remain in the model
  - One could then try to estimate and do inference about $\alpha$ by applying ordinary least squares with $y_i$ as the outcome, and $D_i$ and any selected control variables as regressors.
  - Are there any problems?

# Inference with Selection among Many Controls

- The problem is that it target prediction $\rightarrow$ any variable that is highly correlated to the treatment variable will tend to be dropped
- Of course, the exclusion of a variable that is highly correlated to the treatment will lead to substantial omitted-variables bias

OVB

- It ignores a key component to understanding omitted-variables bias, the relationship between the treatment variable and the controls.

# Inference with Selection among Many Controls

- The naive approach is based on a "structural" model where the target is to learn the treatment effect given controls, not an equation representing a prediction rule for $y_i$ given $D_i$ and $X_i$.

- Let's look it this way

$$\boxed{D_i = X_i'\theta_d + r_{di} + v_i} \tag{12}$$

FWL

  - where $E[v_i|X_i, r_{di}] = 0$
  - but some $\theta_d \neq 0$

- The model we are interested is:

$$y_i = \alpha D_i + X_i'\theta_y + r_{yi} + \zeta_i \tag{13}$$

# Inference with Selection among Many Controls

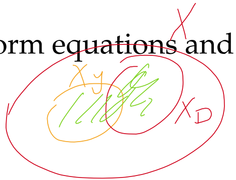
- It is thus useful to transform $y_i = \alpha D_i + X_i'\theta_y + r_{yi} + \zeta_i$ to a reduced form (recall $D_i = X_i'\theta_d + r_{di} + v_i$):

$$y_i = X_i'(\alpha\theta_d + \theta_y) + (\alpha r_{di} + r_{yi}) + r_{di} + (\alpha v_i + \zeta_i) = X_i'\pi + r_{ci} + \epsilon_i \tag{14}$$

- where $E(\epsilon_i|x_i, r_{ci}] = 0$
- $r_{ci}$ is a composite approximation error
- this equation now represent a predictive relationship, which may be estimated using high-dimensional methods.

# Inference with Selection among Many Controls

- To prevent model selection mistakes, it is important to consider both equations for selection.
- We apply variable selection methods to each of the two reduced form equations and then use all of the selected controls in estimation of $\alpha$.

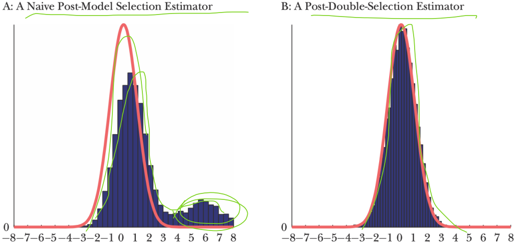



- We select
  1. A set of variables that are useful for predicting $y_i$, say $X_{yi}$, and
  2. A set of variables that are useful for predicting $D_i$, say $X_{di}$.
- We then estimate $\alpha$ by ordinary least squares regression of $y_i$ on $D_i$ and the union of the variables selected for predicting $y_i$ and $D_i$, contained in $X_{yi}$ and $X_{di}$.

$$y_i = \alpha D_i + X_y \theta_y + X_d \theta_d + \zeta_i$$

# Inference with Selection among Many Controls

*Figure 1*
**The "Double Selection" Approach to Estimation and Inference versus a Naive Approach: A Simulation from Belloni, Chernozhukov, and Hansen (forthcoming)**
*(distributions of estimators from each approach)*




*Source:* Belloni, Chernozhukov, and Hansen (forthcoming).
*Notes:* The left panel shows the sampling distribution of the estimator of $\alpha$ 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 $\alpha$ 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.

- We are making sure that we use variables that are important for either of the two predictive relationships to guard against OVB

# Application: Estimation of the treatment effect in a linear model with many confounding factors

- What is the effect of an initial (lagged) level of GDP per capita on the growth rates of GDP per capita?
- Solow-Swan-Ramsey growth model predicts convergence
- Poorer countries should typically grow faster and therefore should tend to catch up with the richer countries, conditional on a set of institutional and societal characteristics.
- Covariates that describe such characteristics include variables measuring education and science policies, strength of market institutions, trade openness, savings rates and others.

# Application: Estimation of the treatment effect in a linear model with many confounding factors

Thus, we are interested in a specification of the form:

$$y_i = \alpha d_i + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{15}$$

where

- $y_i$ is the growth rate of GDP over a specified decade in country $i$,
- $d_i$ is the log of the initial level of GDP at the beginning of the specified period,
- $x_{ij}$'s form a long list of country $i$'s characteristics at the beginning of the specified period.
- We are interested in testing the hypothesis of convergence, $\alpha < 0$.

# Application: Estimation of the treatment effect in a linear model with many confounding factors

For this exercise we use the Barro and Lee (1994) data

```
require("hdm") #package
data(GrowthData) #load data
dim(GrowthData)
```

```
## [1] 90 63
```

The number of covariates $p$ is large relative to the sample size $n$

```
y = GrowthData[,1,drop=F]
d = GrowthData[,3, drop=F]
X = as.matrix(GrowthData)[,-c(1,2,3)]
varnames = colnames(GrowthData)
```

# Application: Estimation of the treatment effect in a linear model with many confounding factors

- Now we can estimate the effect of the initial GDP level.
- First, we estimate by OLS:

```
xnames= varnames[-c(1,2,3)] # names of X variables
dandxnames= varnames[-c(1,2)] # names of D and X variables

# create formulas by pasting names (this saves typing times)
fmla= as.formula(paste("Outcome ~ ", paste(dandxnames, collapse= "+")))

# Estimate using OLS
ls.effect= lm(fmla, data=GrowthData)
```

$y \sim X_1 + X_2 + X_3, \ldots, X_7$

# Application: Estimation of the treatment effect in a linear model with many confounding factors

Second, we estimate the effect by the partialling out by Post-Lasso:

```
dX = as.matrix(cbind(d,X))
lasso.effect = rlassoEffect(x=X, y=y, d=d, method="partialling out")
summary(lasso.effect)
```

FWL

```
## [1] "Estimates and significance testing of the effect of target variables"
##      Estimate. Std. Error t value Pr(>|t|)
## [1,]  -0.04981    0.01394  -3.574 0.000351 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\tilde{y} = \tilde{d}\beta + \varepsilon$

partialled out

FWL

# Application: Estimation of the treatment effect in a linear model with many confounding factors

Third, we estimate the effect by the double selection method:

```
dX = as.matrix(cbind(d,X))
doublesel.effect = rlassoEffect(x=X, y=y, d=d, method="double selection")
summary(doublesel.effect)
```

$y = \alpha D + \underline{X_y} \theta_y + X_d \theta_d + \varepsilon$

```
## [1] "Estimates and significance testing of the effect of target variables"
##          Estimate. Std. Error t value Pr(>|t|)
## gdpsh465  -0.05001    0.01579  -3.167  0.00154 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Application: Estimation of the treatment effect in a linear model with many confounding factors

- Collecting the results

| | Estimate | Std. Error |
|---|---|---|
| full reg via ols | -0.01 | 0.02989 |
| partial reg via post-lasso | -0.05 | 0.01394 |
| partial reg via double selection | -0.05 | 0.01579 |

equivalent

# Review & Next Steps

3 30

- Today:
- Elastic Net
- Lasso for Causality: Post Lasso Double Selection
- Next class: Classification

# Further Readings


- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2), 608-650.
- 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.
- Chernozhukov, V., Hansen, C., & Spindler, M (2016). hdm: High-Dimensional Metrics R Journal, 8(2), 185-199. https://journal.r-project.org/archive/2016/RJ-2016-040/index.html
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal StatisticalSociety, Series B.67: pp. 301–320