09 - High-Dimensional Confounding Adjustment

ml4econ, HUJI 2021

Itamar Caspi
May 30, 2021 (updated: 2021-05-30)

Replicating this presentation

Use the pacman package to install and load packages:

```r
if (!require("pacman"))
  install.packages("pacman")

pacman::p_load(
  tidyverse,
  tidymodels,
  hdm,
  ggdag,
  knitr,
  xaringan,
)
```

Outline

- Lasso and Variable Selection
- High Dimensional Confoundedness
- Empirical Illustration using `hdm`

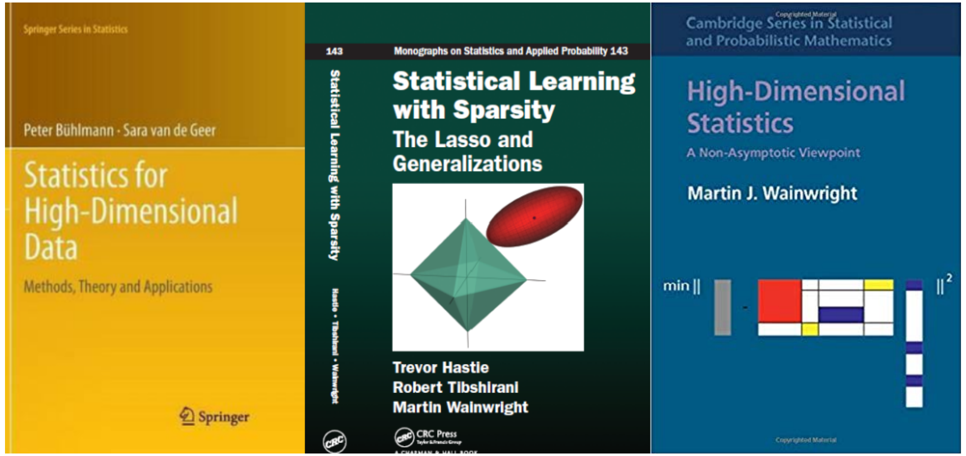
Lasso and Variable Selection

Resources on the theory of Lasso

- *Statistical Learning with Sparsity - The Lasso and Generalizations* (Hastie, Tibshirani, and Wainwright), **Chapter 11: Theoretical Results for the Lasso.** (PDF available online)

- *Statistics for High-Dimensional Data - Methods, Theory and Applications* (Buhlmann and van de Geer), **Chapter 7: Variable Selection with the Lasso.**

- *High Dimensional Statistics - A Non-Asymptotic Viewpoint* (Wainwright), **Chapter 7: Sparse Linear Models in High Dimensions**


Springer Series in Statistics
Peter Bühlmann · Sara van de Geer
Statistics for High-Dimensional Data
Methods, Theory and Applications
Springer
143
Monographs on Statistics and Applied Probability 143
Statistical Learning with Sparsity
The Lasso and Generalizations
Trevor Hastie
Robert Tibshirani
Martin Wainwright
CRC Press
Cambridge Series in Statistical and Probabilistic Mathematics
High-Dimensional Statistics
A Non-Asymptotic Viewpoint
Martin J. Wainwright
min ||
||²

Guarantees vs. guidance

- Most (if not all) of what we've done so far is based on *guidance*
    - Choosing the number of folds in CV
    - Size of the holdout set
    - Tuning parameter(s)
    - loss function
    - function class
- In causal inference, we need *guaranties*
    - variable selection
    - Confidence intervals and p-values
- To get guarantees, we typically need
    - Assumptions about a "true" model
    - Asymptotics $n \to \infty$, $k \to ?$

Some notation to help you penetrate the Lasso literature

Suppose $\boldsymbol{\beta}$ is a $\boldsymbol{k} \times \mathbf{1}$ vector with typical element β_i.

- The ℓ_0-norm is defined as $||\boldsymbol{\beta}||_0 = \sum_{j=1}^{k} \mathbf{1}_{\{\beta_j \neq 0\}}$, i.e., the number of non-zero elements in $\boldsymbol{\beta}$.
- The ℓ_1-norm is defined as $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{k} |\beta_j|$.
- The ℓ_2-norm is defined as $||\boldsymbol{\beta}||_2 = \left(\sum_{j=1}^{k} |\beta_j|^2\right)^{\frac{1}{2}}$, i.e., Euclidean norm.
- The ℓ_∞-norm is defined as $||\boldsymbol{\beta}||_\infty = \sup_j |\beta_j|$, i.e., the maximum entries' magnitude of $\boldsymbol{\beta}$.
- The support of $\boldsymbol{\beta}$, is defined as $S \equiv \text{supp}(\boldsymbol{\beta}) = \{\beta_j \neq 0, j = 1, \ldots, j\}$, i.e., the subset of non-zero coefficients.
- The size of the support $s = |S|$ is the number of non-zero elements in $\boldsymbol{\beta}$, i.e., $s = ||\boldsymbol{\beta}||_0$

Lasso: The basic setup

The linear regression model:

$$Y_i = \alpha + X_i' \boldsymbol{\beta}^0 + \varepsilon_i, \quad i = 1, \ldots, n,$$

$$\mathbb{E}\left[\varepsilon_i X_i\right] = 0, \quad \alpha \in \mathbb{R}, \quad \boldsymbol{\beta}^0 \in \mathbb{R}^k.$$

Under the *exact sparsity* assumption, only a subset of variables of size $\boldsymbol{s} \ll \boldsymbol{k}$ is included in the model where $\boldsymbol{s} \equiv \|\boldsymbol{\beta}\|_0$ is the sparsity index.

$$\underbrace{\mathbf{X}_S = \left(X_{(1)}, \ldots, X_{(s)}\right),}_{\text{sparse variables}} \quad \underbrace{\mathbf{X}_{S^c} = \left(X_{(s+1)}, \ldots, X_{(k)}\right)}_{\text{non-sparse variables}}$$

where S is the subset of active predictors, $\mathbf{X}_S \in \mathbb{R}^{n \times s}$ corresponds to the subset of covariates that are in the sparse set, and $\mathbf{X}_{S^C} \in \mathbb{R}^{n \times k-s}$ is the subset of the "irrelevant" non-sparse variables.

Lasso: Optimization

Lasso (least absolute shrinkage and selection operator) was introduced by Tibshirani (1996). The optimization problem:

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \|\boldsymbol{\beta}\|_1$$

Lasso puts a "budget constraint" on the sum of *absolute* β's.

Unlike ridge, the lasso penalty is linear (moving from 1 to 2 is the same as moving from 101 to 102.)

A great advantage of the lasso is that performs model selection - it zeros out most of the β's in the model (the solution is *sparse*.)

Any penalty that involves the ℓ_1 norm will do this.

Evaluation of the Lasso

Let β^0 denote the true vector of coefficients and let $\widehat{\beta}$ denote the Lasso estimator.

We can asses the quality of the Lasso in several ways:

I. Prediction quality

$$\text{Loss}_{\text{ pred}} \left(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}^0\right) = \frac{1}{N}\left\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\mathbf{X}\right\|_2^2 = \frac{1}{N}\sum_{j=1}^{k}\left[(\hat{\beta}_j - \beta_j^0)\mathbf{X}_{(j)}\right]^2$$

II. Parameter consistency

$$\text{Loss}_{\text{param}} \left(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}^0\right) = \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_2^2 = \sum_{j=1}^{k}(\hat{\beta} - \beta^0)^2$$

III. Support recovery (sparsistency), e.g., $+1$ if $\mathbf{sign}(\beta^0) = \mathbf{sign}(\beta_j)$, for all $\boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{k}$, and zero otherwise.

Lasso as a variable selection tool

- Variable selection consistency is essential for causal inference (think omitted variable bias).

- Lasso is often used as a variable selection tool.

- Being able to select the "true" support by Lasso relies on strong assumptions about

  - the ability to distinguish between relevant and irrelevant variables.
  - the ability to identify $\boldsymbol{\beta}$.

Critical assumption #1: Distinguishable sparse betas

Lower eigenvalue: the min eigenvalue $\lambda_{\min}$ of the sub-matrix $\mathbf{X}_S$ is bounded away from zero.

$$\lambda_{\min}\left(\mathbf{X}_S'\mathbf{X}_S/N\right) \geq C_{\min} > 0$$

Linear dependence between the columns of $\mathbf{X}_s$ would make it impossible to identify the true $\boldsymbol{\beta}$, even if we *knew* which variables are included in $\mathbf{X}_S$.

NOTE: The high-dimension's lower eigenvalue condition replaces the low-dimension's rank condition (i.e., that $\mathbf{X}'\mathbf{X}$ is invertible)

Critical assumption #2: Distinguishable active predictors

Irrepresentability condition (Zou ,2006; Zhao and Yu, 2006): There must exist some $\eta \in [0, 1)$ such that

$$\max_{j \in S^c} \left\| \left(\mathbf{X}_S' \mathbf{X}_S \right)^{-1} \mathbf{X}_S' \mathbf{x}_j \right\|_1 \leq \eta$$

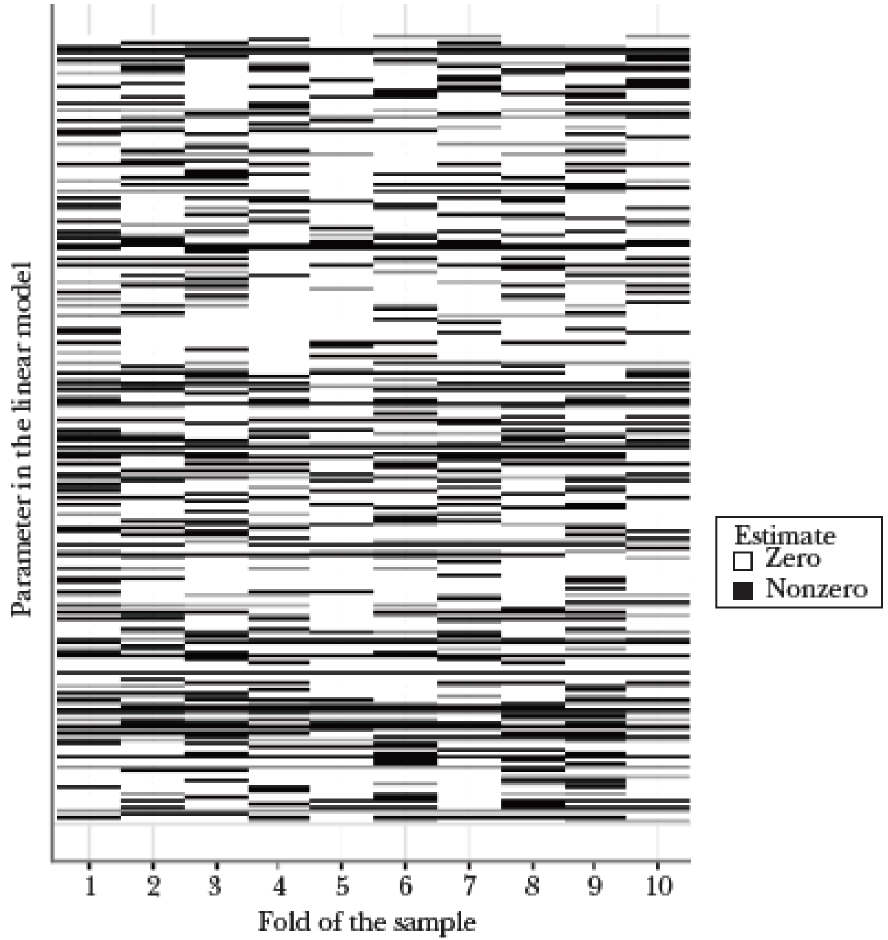
INTUITION: What's inside $\|\cdot\|_1$ is like regressing $\mathbf{x}_j$ on the variables in $\mathbf{X}_s$.

- When $\eta = 0$, the sparse and non-sparse variables are orthogonal to each other.
- When $\eta = 1$, we can reconstruct (some elements of) $\mathbf{X}_S$ using $\mathbf{X}_{S^C}$.

Thus, the irrepresentability condition roughly states that we can distinguish the sparse variables from the non-sparse ones.

Figure 2

Selected Coefficients (Nonzero Estimates) across Ten LASSO Regressions



Source: Mullainathan and Spiess (JEP 2017).

Some words on setting the optimal tuning parameter

- As we've seen thorough this course, it is also common to choose λ empirically, often by cross-validation, based on its predictive performance
- In causal analysis, inference and not prediction is the end goal. Moreover, these two objectives often contradict each other (bias vs. variance)
- Optimally, the choice of λ should provide guarantees about the performance of the model.
- Roughly speaking, when it comes to satisfying sparsistency, λ is set such that it selects non-zero β's with high probability.

High Dimensional Confoundedness

"Naive" implementation of the Lasso

Run `glmnet`

```
glmnet(Y ~ DX)
```

where `DX` is the feature matrix which includes both the treatment D and the features vector X.

The estimated coefficients are:

$$\left(\widehat{\alpha}, \hat{\tau}, \widehat{\boldsymbol{\beta}}'\right)' = \underset{\alpha,\tau\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau D_i - \boldsymbol{\beta}' X_i\right)^2 + \lambda\left(|\tau| + \sum_{j=1}^{k} |\beta_j|\right)$$

PROBLEMS:

1. Both $\hat{\tau}$ and $\widehat{\boldsymbol{\beta}}$ are biased towards zero (shrinkage).
2. Lasso might drop D_i, i.e., shrink $\hat{\tau}$ to zero. Can also happen to relevant confounding factors.
3. How to choose λ?

Toward a solution

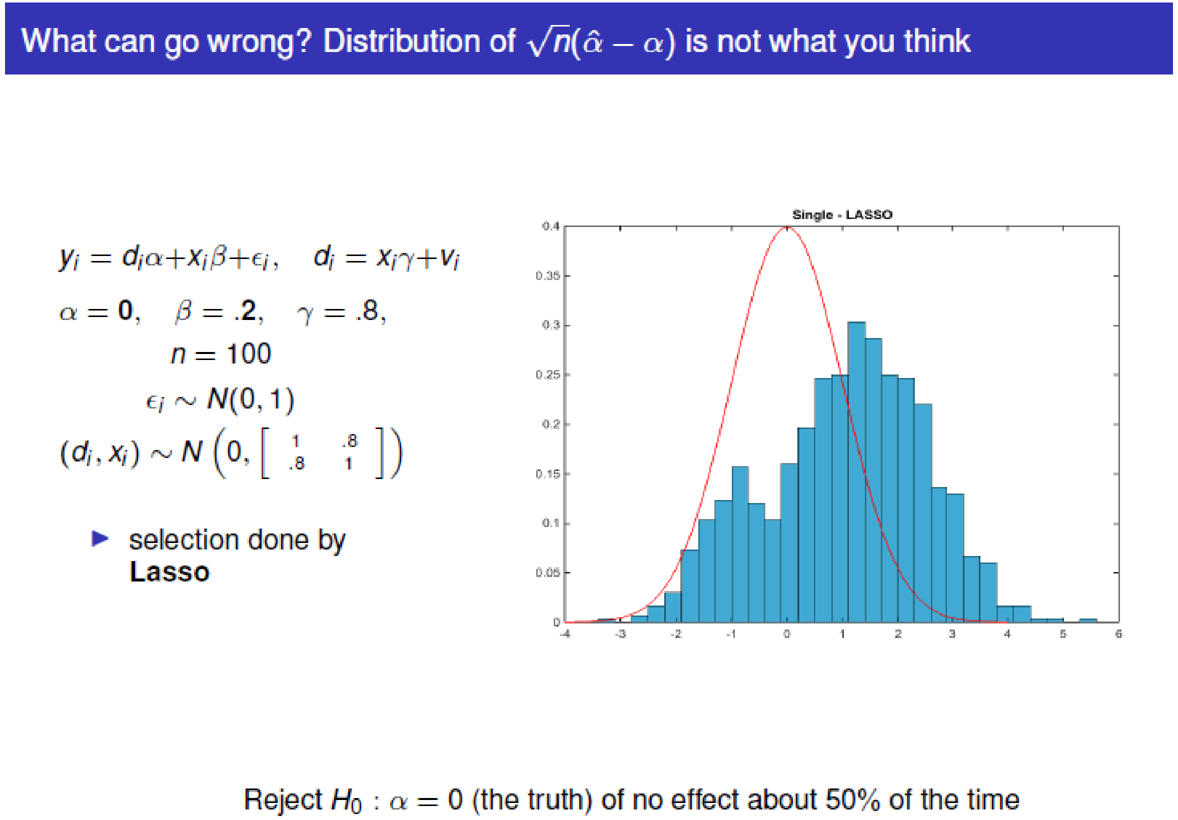
OK, lets keep D_i in:

$$\left(\widehat{\alpha}, \hat{\tau}, \widehat{\boldsymbol{\beta}}'\right)' = \underset{\alpha,\tau\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^k}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau D_i - \boldsymbol{\beta}' X_i\right)^2 + \lambda \left(\sum_{j=1}^{k} |\beta_j|\right)$$

Then, *debias* the results using "Post-Lasso", i.e, use Lasso for variable selection and then run OLS with the selected variables.

PROBLEMS: *Omitted variable bias*. The Lasso might drop features that are correlated with D_i because they are "bad" predictor of Y_i.

Problem solved?



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

Source: https://stuff.mit.edu/~vchern/papers/Chernozhukov-Saloniki.pdf

Solution: Double-selection Lasso (Belloni, et al., REStud 2013)

First step: Regress Y_i on X_i and D_i on X_i:

$$\widehat{\gamma} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{\gamma}' X_i\right)^2 + \lambda_\gamma \left(\sum_{j=2}^{p} |\gamma_j|\right)$$

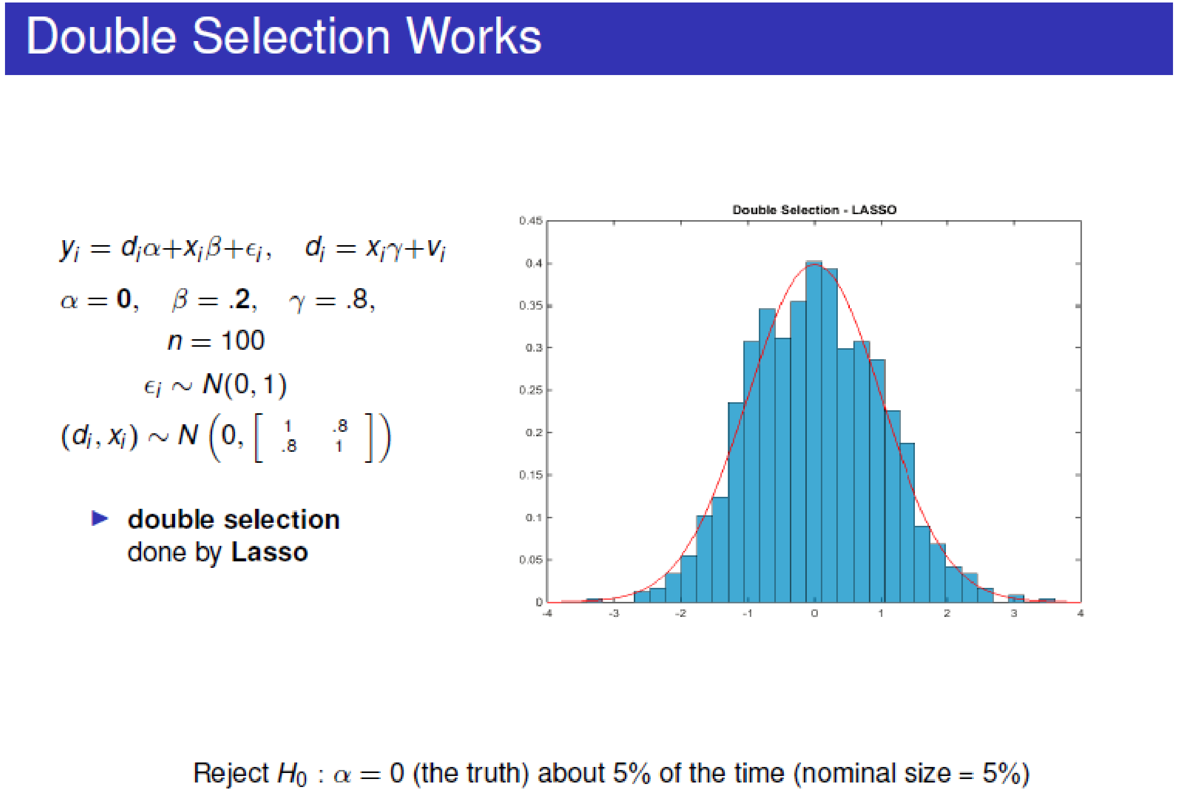
$$\hat{\delta} = \underset{\boldsymbol{\delta} \in \mathbb{R}^{q+1}}{\arg\min} \sum_{i=1}^{n} \left(D_i - \boldsymbol{\delta}' X_i\right)^2 + \lambda_\delta \left(\sum_{j=2}^{q} |\delta_j|\right)$$

Second step: Refit the model by OLS and include the $\mathbf{X}$'s that are significant predictors of Y_i *and* D_i.

Third step: Proceed to inference using standard confidence intervals.

> The Tuning parameter λ is set such that the non-sparse coefficients are correctly selected with high probability.

Does it work?



Double Selection Works

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

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

$$n = 100$$

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

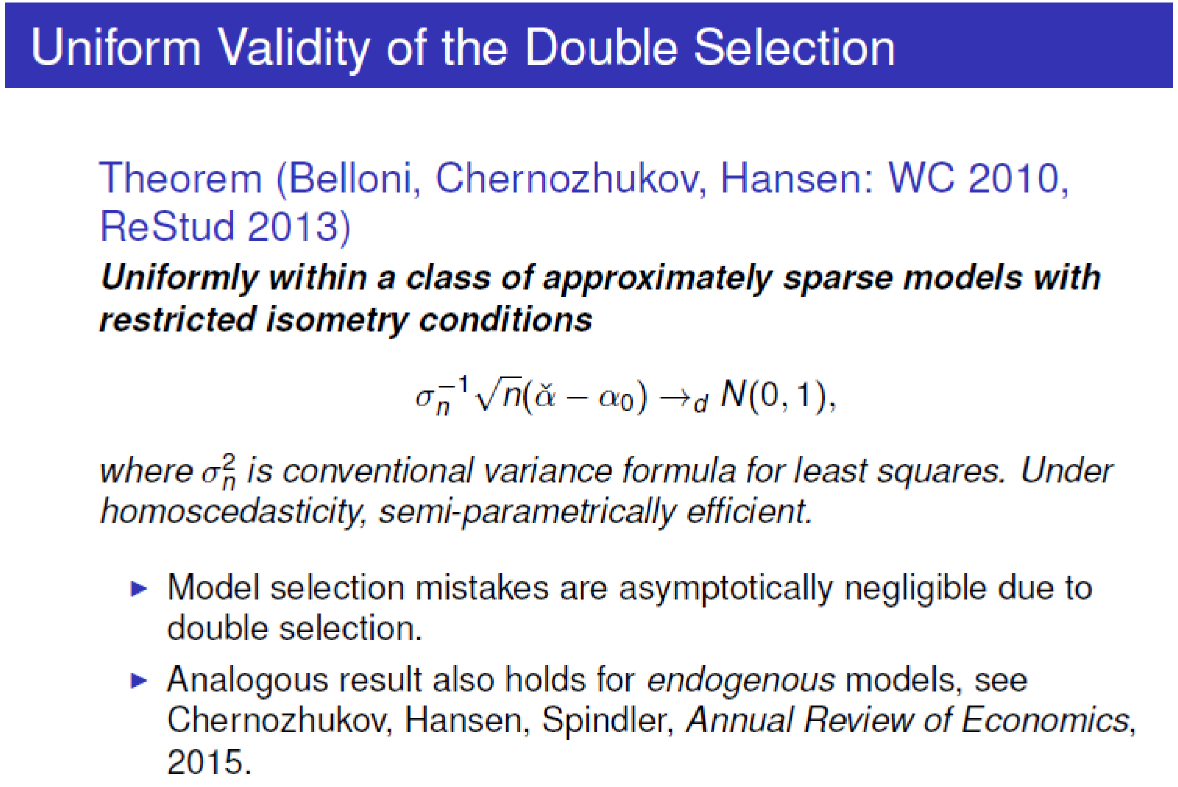
$$(d_i, x_i) \sim N\left(0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}\right)$$

- **double selection** done by **Lasso**

Reject $H_0 : \alpha = 0$ (the truth) about 5% of the time (nominal size = 5%)

Source: https://stuff.mit.edu/~vchern/papers/Chernozhukov-Saloniki.pdf

Statistical inference



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

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

where σ_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.

Source: https://stuff.mit.edu/~vchern/papers/Chernozhukov-Saloniki.pdf

Intuition: Partialling-out regression

consider the following two alternatives for estimating the effect of X_{1i} (a scalar) on Y_i, while adjusting for X_{2i}:

Alternative 1: Run

$$Y_i = \alpha + \beta X_{1i} + \gamma X_{2i} + \varepsilon_i$$

Alternative 2: First, run Y_i on X_{2i} and X_{1i} on X_{2i} and keep the residuals, i.e., run

$$Y_i = \gamma_0 + \gamma_1 X_{2i} + u_i^Y, \quad \text{and} \quad X_{1i} = \delta_0 + \delta_1 X_{2i} + u_i^{X_1},$$

and keep $\widehat{u}_i^Y$ and $\widehat{u}_i^{X_1}$. Next, run

$$\widehat{u}_i^Y = \beta^* \widehat{u}_i^{X_1} + v_i.$$

According to the Frisch-Waugh-Lovell (FWV) Theorem, $\widehat{\beta} = \widehat{\beta}^*$.

Notes on the guarantees of double-selection Lasso

Approximate Sparsity Consider the following regression model:

$$Y_i = f(W_i) + \varepsilon_i = X_i' \boldsymbol{\beta}^0 + r_i + \varepsilon_i, \quad 1, \ldots, n$$

where r_i is the approximation error.

Under *approximate sparsity*, it is assumed that $f(W_i)$ can be approximated sufficiently well (up to r_i) by $X_i' \boldsymbol{\beta}^0$, while using only a small number of non-zero coefficients.

Restricted Sparse Eigenvalue Condition (RSEC) This condition puts bounds on the number of variables outside the support the Lasso can select. Relevant for the post-lasso stage.

Regularization Event The tuning parameter λ is to a value that it selects to correct model with probability of at least p, where p is set by the user. Further assumptions regarding the quantile function of the maximal value of the gradient of the objective function at $\boldsymbol{\beta}^0$, and the error term (homoskedasticity vs. heteroskedasticity). See Belloni et al. (2012) for further details.

Further extensions of double-selection

1. Chernozhukov et al. (AER 2017): Other function classes ("Double-ML"), e.g., use random forest for $Y \sim X$ and regularized logit for $D \sim X$.

2. Instrumental variables (Belloni et al., Ecta 2012, Chernozhukov et al., AER 2015), see **problem set**.

3. Heterogeneous treatment effects (Belloni et al., Ecta 2017), **next week**.

4. Panel data (Belloni, et al., JBES 2016)

Recent evidence on the applicability of double-lasso

"Machine Labor" (Angrist and Frandsen, 2019):

- Application of double lasso to estimation of effects of elite college attendance (Dale and Kruger, 2002) shows that the resulting estimates of causal effects are stable, consistently showing little evidence of an elite college advantage
- The authors' findings on double lasso (and double ML in general) in IV applications (Angrist and Krueger, 1991) are less encouraging: double lasso IV screening sometimes outperform OLS. But, standard non-ML methods (e.g., LIML) do better.

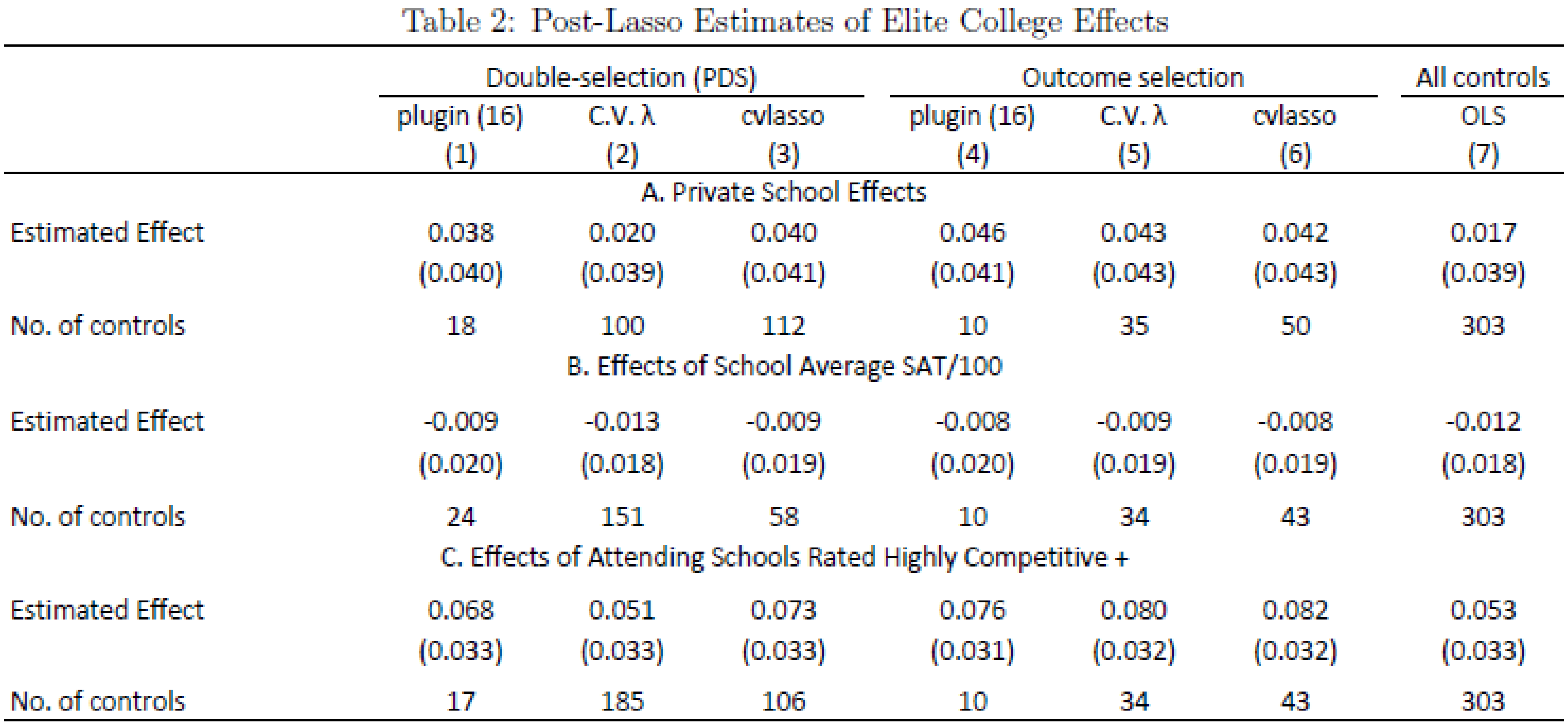


Table 2: Post-Lasso Estimates of Elite College Effects

	Double-selection (PDS)			Outcome selection			All controls
	plugin (16)	C.V. λ	cvlasso	plugin (16)	C.V. λ	cvlasso	OLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Private School Effects							
Estimated Effect	0.038	0.020	0.040	0.046	0.043	0.042	0.017
	(0.040)	(0.039)	(0.041)	(0.041)	(0.043)	(0.043)	(0.039)
No. of controls	18	100	112	10	35	50	303
B. Effects of School Average SAT/100							
Estimated Effect	-0.009	-0.013	-0.009	-0.008	-0.009	-0.008	-0.012
	(0.020)	(0.018)	(0.019)	(0.020)	(0.019)	(0.019)	(0.018)
No. of controls	24	151	58	10	34	43	303
C. Effects of Attending Schools Rated Highly Competitive +							
Estimated Effect	0.068	0.051	0.073	0.076	0.080	0.082	0.053
	(0.033)	(0.033)	(0.033)	(0.031)	(0.032)	(0.032)	(0.033)
No. of controls	17	185	106	10	34	43	303

Source: Angrist and Frandsen (2019).

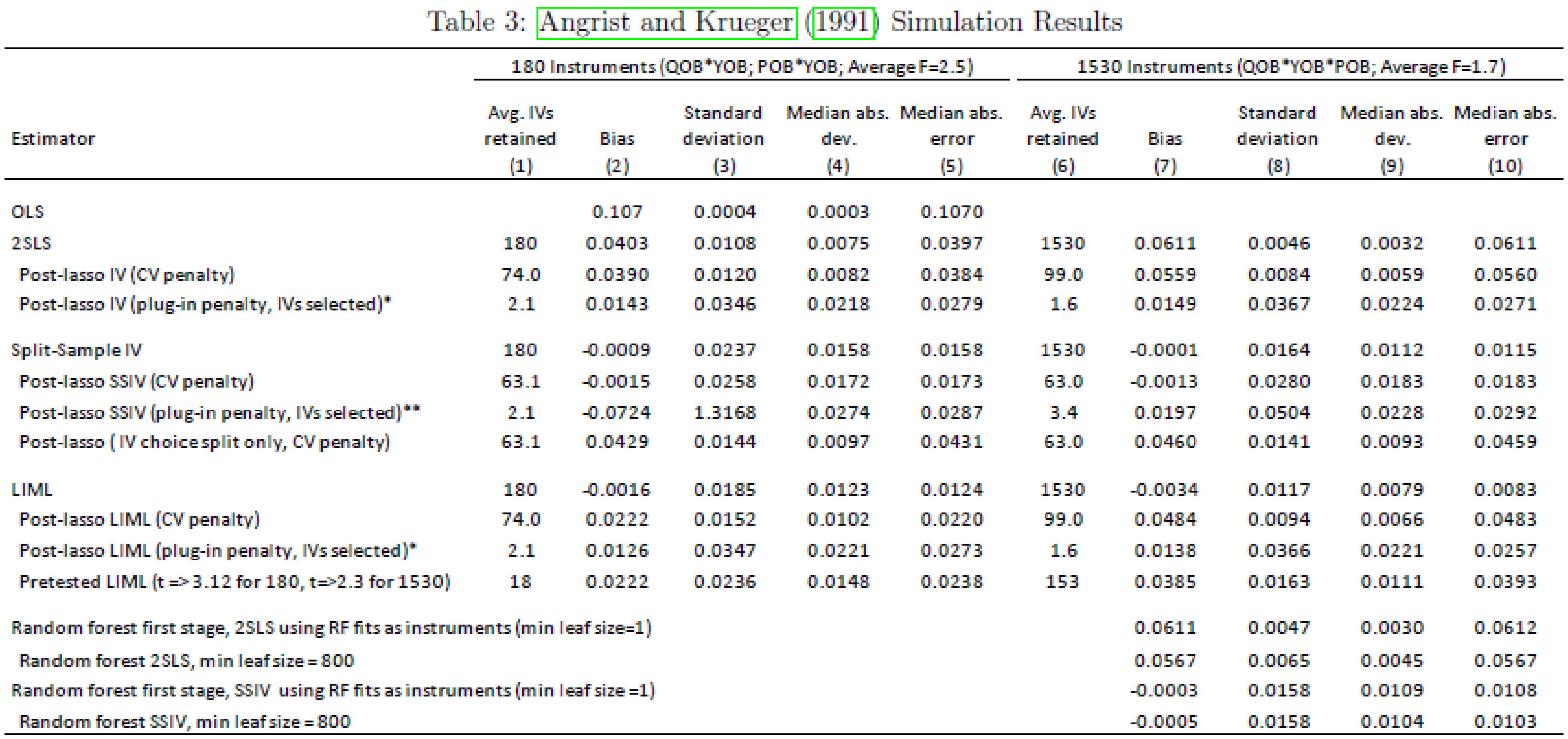


Table 3: Angrist and Krueger (1991) Simulation Results

Estimator	180 Instruments (QOB*YOB; POB*YOB; Average F=2.5)					1530 Instruments (QOB*YOB*POB; Average F=1.7)				
	Avg. IVs retained	Bias	Standard deviation	Median abs. dev.	Median abs. error	Avg. IVs retained	Bias	Standard deviation	Median abs. dev.	Median abs. error
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
OLS		0.107	0.0004	0.0003	0.1070					
2SLS	180	0.0403	0.0108	0.0075	0.0397	1530	0.0611	0.0046	0.0032	0.0611
Post-lasso IV (CV penalty)	74.0	0.0390	0.0120	0.0082	0.0384	99.0	0.0559	0.0084	0.0059	0.0560
Post-lasso IV (plug-in penalty, IVs selected)*	2.1	0.0143	0.0346	0.0218	0.0279	1.6	0.0149	0.0367	0.0224	0.0271
Split-Sample IV	180	-0.0009	0.0237	0.0158	0.0158	1530	-0.0001	0.0164	0.0112	0.0115
Post-lasso SSIV (CV penalty)	63.1	-0.0015	0.0258	0.0172	0.0173	63.0	-0.0013	0.0280	0.0183	0.0183
Post-lasso SSIV (plug-in penalty, IVs selected)**	2.1	-0.0724	1.3168	0.0274	0.0287	3.4	0.0197	0.0504	0.0228	0.0292
Post-lasso (IV choice split only, CV penalty)	63.1	0.0429	0.0144	0.0097	0.0431	63.0	0.0460	0.0141	0.0093	0.0459
LIML	180	-0.0016	0.0185	0.0123	0.0124	1530	-0.0034	0.0117	0.0079	0.0083
Post-lasso LIML (CV penalty)	74.0	0.0222	0.0152	0.0102	0.0220	99.0	0.0484	0.0094	0.0066	0.0483
Post-lasso LIML (plug-in penalty, IVs selected)*	2.1	0.0126	0.0347	0.0221	0.0273	1.6	0.0138	0.0366	0.0221	0.0257
Pretested LIML (t => 3.12 for 180, t=>2.3 for 1530)	18	0.0222	0.0236	0.0148	0.0238	153	0.0385	0.0163	0.0111	0.0393
Random forest first stage, 2SLS using RF fits as instruments (min leaf size=1)							0.0611	0.0047	0.0030	0.0612
Random forest 2SLS, min leaf size = 800							0.0567	0.0065	0.0045	0.0567
Random forest first stage, SSIV using RF fits as instruments (min leaf size =1)							-0.0003	0.0158	0.0109	0.0108
Random forest SSIV, min leaf size = 800							-0.0005	0.0158	0.0104	0.0103

Source: Angrist and Frandsen (2019).

Empirical Illustration using `hdm`

The hdm package*

"High-Dimensional Metrics" (`hdm`) by Victor Chernozhukov, Chris Hansen, and Martin Spindler is an R package for estimation and quantification of uncertainty in high-dimensional approximately sparse models.

[*] There is also a rather new Stata module named Lassopack that includes a rich suite of programs for regularized regression in high-dimensional setting.

Illustration: Testing for growth convergence

The standard growth convergence empirical model:

$$Y_{i,T} = \alpha_0 + \alpha_1 Y_{i,0} + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where

- $Y_{i,T}$ national growth rates in GDP per capita for the periods 1965-1975 and 1975-1985.
- $Y_{i,0}$ is the log of the initial level of GDP at the beginning of the specified decade.
- X_{ij} covariates which might influence growth.

The growth convergence hypothesis implies that $\alpha_1 < 0$.

Growth data

To test the growth convergence hypothesis, we will make use of the Barro and Lee (1994) dataset

```
data("GrowthData")
```

The data contain macroeconomic information for large set of countries over several decades. In particular,

- n = 90 countries
- k = 60 country features

Not so big...

Nevertheless, the number of covariates is large relative to the sample size $\Rightarrow$ variable selection is important!

```r
GrowthData %>%
  as_tibble %>%
  head(2)
```

```
## # A tibble: 2 x 63
##   Outcome intercept gdpsh465 bmp1l freeop freetar   h65  hm65  hf65   p65  pm65
##     <dbl>     <int>    <dbl> <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 -0.0243         1     6.59 0.284  0.153  0.0439 0.007 0.013 0.001  0.29  0.37
## 2  0.100          1     6.83 0.614  0.314  0.0618 0.019 0.032 0.007  0.91  1
## # ... with 52 more variables: pf65 <dbl>, s65 <dbl>, sm65 <dbl>, sf65 <dbl>,
## #   fert65 <dbl>, mort65 <dbl>, lifee065 <dbl>, gpop1 <dbl>, fert1 <dbl>,
## #   mort1 <dbl>, invsh41 <dbl>, geetot1 <dbl>, geerec1 <dbl>, gde1 <dbl>,
## #   govwb1 <dbl>, govsh41 <dbl>, gvxdxe41 <dbl>, high65 <dbl>, highm65 <dbl>,
## #   highf65 <dbl>, highc65 <dbl>, highcm65 <dbl>, highcf65 <dbl>, human65 <dbl>,
## #   humanm65 <dbl>, humanf65 <dbl>, hyr65 <dbl>, hyrm65 <dbl>, hyrf65 <dbl>,
## #   no65 <dbl>, nom65 <dbl>, nof65 <dbl>, pinstab1 <dbl>, pop65 <int>,
## #   worker65 <dbl>, pop1565 <dbl>, pop6565 <dbl>, sec65 <dbl>, secm65 <dbl>,
## #   secf65 <dbl>, secc65 <dbl>, seccm65 <dbl>, seccf65 <dbl>, syr65 <dbl>,
## #   syrm65 <dbl>, syrf65 <dbl>, teapri65 <dbl>, teasec65 <dbl>, ex1 <dbl>,
## #   im1 <dbl>, xr65 <dbl>, tot1 <dbl>
```

Data processing

Rename the response and "treatment" variables:

```
df <-
  GrowthData %>%
  rename(YT = Outcome, Y0 = gdpsh465)
```

Transform the data to vectors and matrices (to be used in the `rlassoEffect()` function)

```
YT <- df %>% select(YT) %>% pull()

Y0 <- df %>% select(Y0) %>% pull()

X <- df %>%
  select(-c("Y0", "YT")) %>%
  as.matrix()

Y0_X <- df %>%
  select(-YT) %>%
  as.matrix()
```

Estimation of the convergence parameter α_1

Method 1: OLS

```
ols <- lm(YT ~ ., data = df)
```

Method 2: Naive (rigorous) Lasso

```
naive_Lasso <- rlasso(x = Y0_X, y = YT)
```

Does the Lasso drop `Y0`?

```
naive_Lasso$beta[2]
```

```
## Y0
##  0
```

Unfortunately, yes...

Estimation of the convergence parameter α_1

Method 3: Partialling out Lasso

```
part_Lasso <-
  rlassoEffect(
    x = X, y = YT, d = Y0,
    method = "partialling out"
  )
```

Method 4: Double-selection Lasso

```
double_Lasso <-
  rlassoEffect(
    x = X, y = YT, d = Y0,
    method = "double selection"
  )
```

Tidying the results

```r
# OLS
ols_tbl <- tidy(ols) %>%
  filter(term == "Y0") %>%
  mutate(method = "OLS") %>%
  select(method, estimate, std.error)

# Naive Lasso
naive_Lasso_tbl <- tibble(method = "Naive Lasso",
                              estimate = NA,
                              std.error = NA)
# Partialling-out Lasso
results_part_Lasso <- summary(part_Lasso)[[1]][1, 1:2]
part_Lasso_tbl     <- tibble(method = "Partialling-out Lasso",
                          estimate = results_part_Lasso[1],
                          std.error = results_part_Lasso[2])
# Double-selection Lasso
results_double_Lasso <- summary(double_Lasso)[[1]][1, 1:2]
double_Lasso_tbl <- tibble(method = "Double-selection Lasso",
                           estimate = results_double_Lasso[1],
                           std.error = results_double_Lasso[2])
```

Results of the convergence test

```r
bind_rows(ols_tbl, naive_Lasso_tbl, part_Lasso_tbl, double_Lasso_tbl) %>%
  kable(digits = 3, format = "html")
```

method	estimate	std.error
OLS	-0.009	0.030
Naive Lasso	NA	NA
Partialling-out Lasso	-0.050	0.014
Double-selection Lasso	-0.050	0.016

Double-selection and partialling-out yield much more precise estimates and provide support the conditional convergence hypothesis

Another (more advanced) R Package

- The Python and R packages {DoubleML} provide an up-to-date implementation of the double / debiased machine learning framework of Chernozhukov et al. (2018).
- See the Getting Started and Examples sections for more information.
- The package is built upon the {mlr3} ecosystem.

`slides %>% end()`

Source code

Selected references

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2019). lassopack: Model selection and prediction with regularized regression in Stata.

Angrist, Joshua D, and Alan B Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106(4): 979–1014.

Angrist, J., & Frandsen, B. (2019). Machine Labor (No. w26584). National Bureau of Economic Research.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica* 80(6): 2369–2429.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547.

Belloni, A., Chernozhukov, V., & Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608–650.

Selected references

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. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5), 486–490.

Chernozhukov, V., Hansen, C., & Spindler, M. (2016). hdm: High-Dimensional Metrics. *The R Journal*, 8(2), 185–199.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5), 261–265.

Selected references

Dale, Stacy Berg, and Alan B Krueger. 2002. "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables." *The Quarterly Journal of Economics*, 117(4): 1491–1527.

Mullainathan, S. & Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87–106.

Van de Geer, S. A., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3, 1360–1392.

Zhao, P., & Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541–2563.