Labour Economics and Big Data: An Overview of Regularization Techniques in the Evaluation of Causal Effects

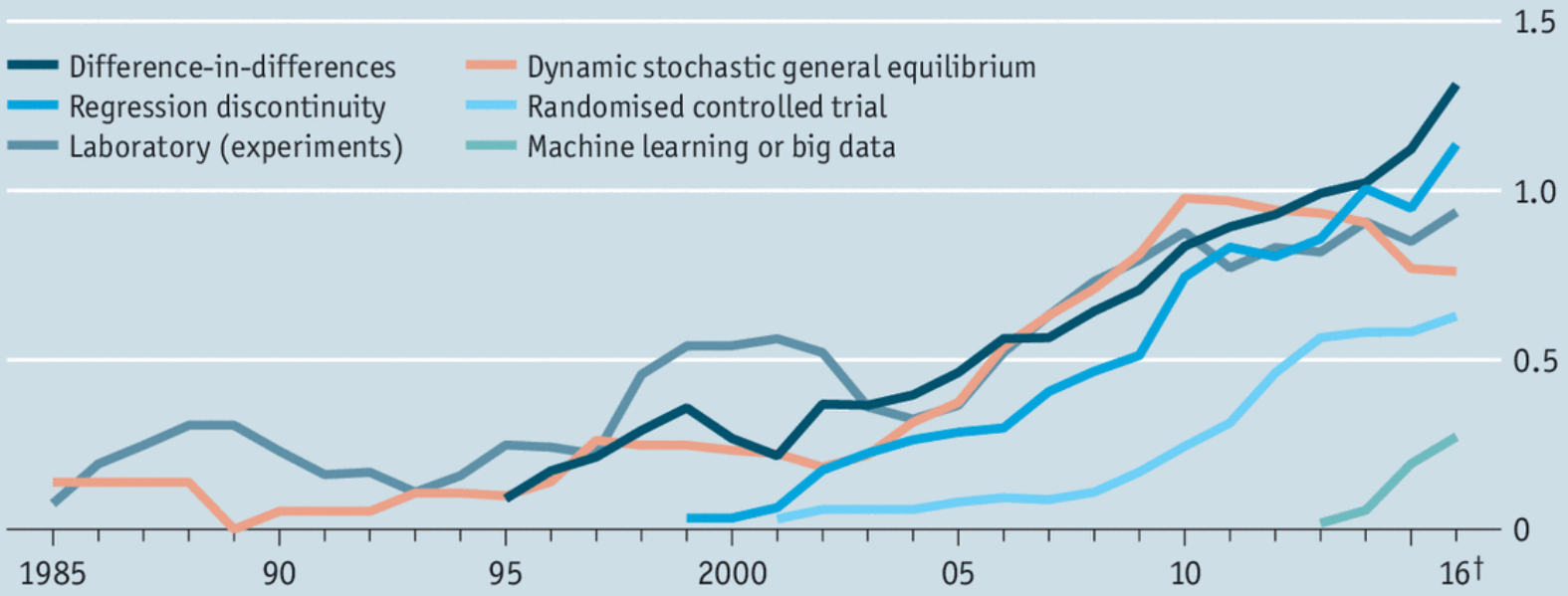
Juan J. Dolado (UC3M)

Conferencia FUNCAS: ***Nuevos Enfoques de Problemas Económicos con Big Data***

Race against the machine (learning)

Dedicated followers of fashion

Mentions in NBER working-paper abstracts, % of total papers*

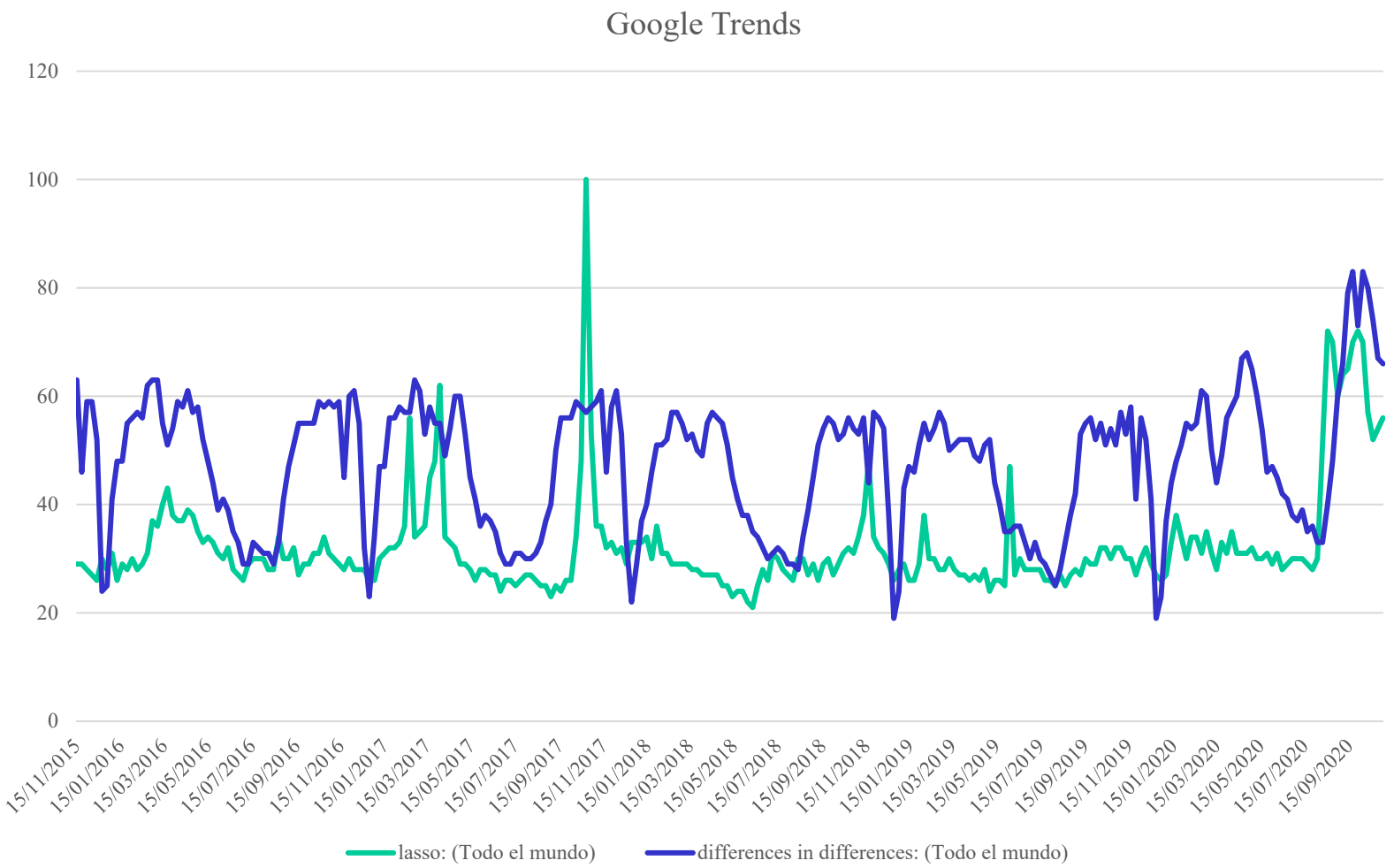


Sources: NBER; *The Economist*

*Five-year moving average †To November

Economist.com

Catching up


Google Trends
120
100
80
60
40
20
0
15/11/2015
15/01/2016
15/03/2016
15/05/2016
15/07/2016
15/09/2016
15/11/2016
15/01/2017
15/03/2017
15/05/2017
15/07/2017
15/09/2017
15/11/2017
15/01/2018
15/03/2018
15/05/2018
15/07/2018
15/09/2018
15/11/2018
15/01/2019
15/03/2019
15/05/2019
15/07/2019
15/09/2019
15/11/2019
15/01/2020
15/03/2020
15/05/2020
15/07/2020
15/09/2020
lasso: (Todo el mundo)
differences in differences: (Todo el mundo)

Outline

- Are Machine Learning (ML) techniques interesting at all for labour economists (LE)? $\hat{y}$ vs. $\hat{\beta}$
- Estimation of causal effects: OLS & IV
- Post Selection (PS) vs. Double Post-Selection in

  (a) Linear Models with exogenous covariates

  (b) Linear Models with endogenous treatment

  (c) Nonlinear models

- Empirical application: AK (1991)

Why ML for LE?

- **ML is making a growing impact on economics with the advent of Big Data (administrative data)**

Belloni et al (2104), Mullanaithan & Spiess (2017), Athey and Imbens (2019), Angrist & Brigham (2020).

- **ML procedures (Lasso, Random trees, Random forests, Bagging, Boosting, Neural networks) aim at improving forecasts**

- **In Labour Economics (LE) we care about β rather than $\hat{y}$ or $R^2 s$ in experimental contexts**

- **But in non-randomized setups, need to control for selection on observables (OLS, IV) and their dimension may be very high**

  $\rightarrow$ need to regularize # controls, instruments in 1st-stage

Standard model in LE

$d_i :=$ treatment, $\quad i = 1,2,\dots,n$

$$
\begin{aligned}
y_{0i} &= \mu + v_i, \\
y_{1i} &= \alpha + y_{0ii} \\
y_i &= \mu + \alpha d_i + v_i,
\end{aligned}
$$

$$\mathbb{E}(v_i \mid d_i = 1, \mathbf{x}_i = \mathbf{x}) = \mathbb{E}(v_i \mid d_i = 0, \mathbf{x}_i = \mathbf{x}) \qquad \textbf{CMI}$$

$$y_i = \mu + \alpha d_i + \beta' \mathbf{x}_i + \varepsilon_i, \qquad \textbf{OLS}$$

$$
\begin{aligned}
y_i &= \alpha d_i + \beta' \mathbf{x}_i + \varepsilon_i, \\
d_i &= \gamma' \mathbf{z}_i + e_i, \\
\mathbf{z}_i &= \varphi' \mathbf{x}_i + \omega_i,
\end{aligned}
\qquad \textbf{IV}
$$

Selection in linear models with exogenous variables

$$y_i = \alpha d_i + \sum_{j=1}^{p} \beta_j \, x_{ij} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid \mathbf{x}_i, z_i) = 0$$

with p≅ n or p>>n

Post selection (**PS**):

(i) Use ML (e.g. Lasso) to select x_{ij} if it´s a relevant predictor for y_i (excluding d_i)

(i) Estimate α by OLS using selected $x´s$ and d_i

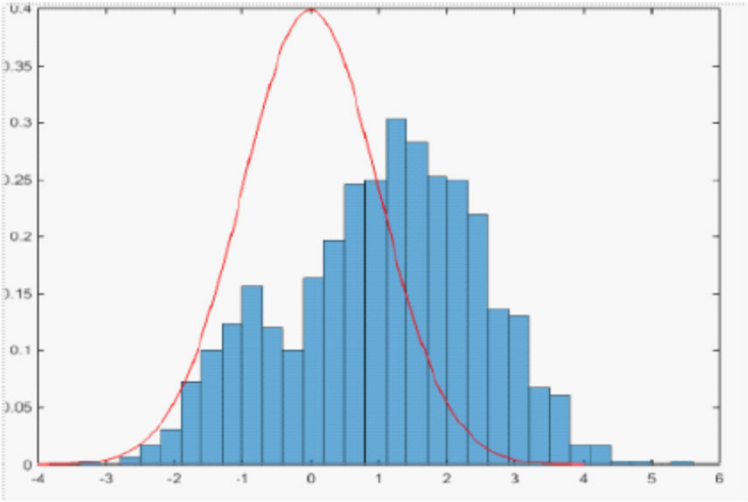
The perils of PS (t-ratio or Lasso with *p*=1)

DGP

$$y_i = \alpha d_i + \beta x_i + \varepsilon_i,$$
$$d_i = \gamma x_i + \nu_i$$

$$\alpha = 0, \beta = 0.2, \gamma = 0.8, \varepsilon \sim N(0,1), v \sim N(0,.32), E(\varepsilon v) = 0,$$
$$x_i \sim N(0,1), \quad n = 100, \#simulations = 1000$$

Figura 1 (Post Selección (PS): t-ratio al 5%)



Why?: reduced form

$$y_i = (\alpha\gamma + \beta)\, x_i + (\varepsilon_i + \alpha\nu_i) = \pi_i\, x_i + \eta_i$$

π is small & γ high !!

Similar with Lasso: effective size (47%)>> nominal size (5%)

Double Post selection (**DPS)**

(i) (i) Use ML (e.g. Lasso) to select x_{ij} if it´s a relevant regressor for y_i and d_i

(ii) Estimate α by OLS using the union of selected x´s and d_i (or residualise: Frisch-Waugh- Lovell Theorem)

Assumptions (Belloni et al., 2013, 2014,…)

- *Approximate Sparsity*: only a subset of regressors $s_n << p_n$ is relevant. E.g. $|\beta_j| < Aj^{-a},\ a > 1.$

- *Restricted Isometry*: limited dependence among subsets of regressors

Comparison with previous DGP (now $p>>n$)

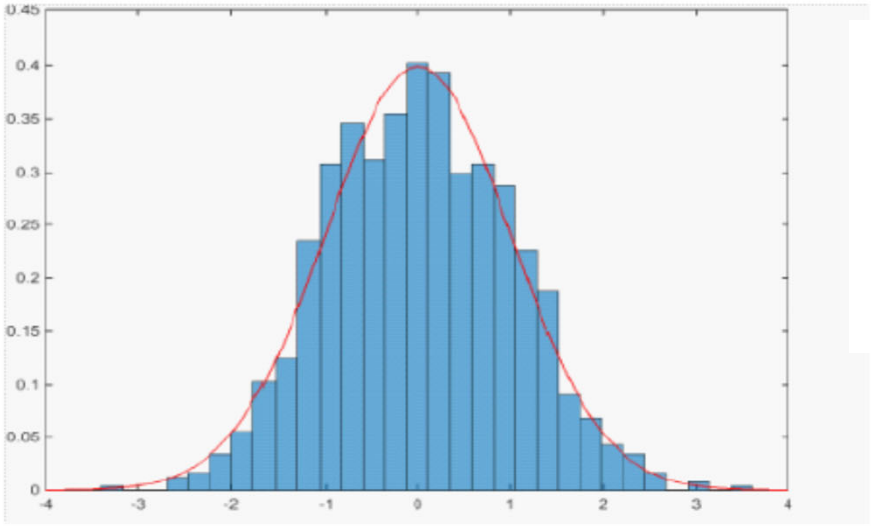
DGP

$$y_i = \alpha d_i + \sum_{j=1}^{p} \beta_i \, x_{ij} + \varepsilon_i,$$

$$d_i = \sum_{j=1}^{p} \gamma_j \, x_{ij} + \nu_i,$$

With p=200, n=100, $\beta_j \& \gamma_j \approx O(j^{-2})$, $H_0: \alpha = 0$

Figura 2 (Post Selección Doble (PSD): Lasso al 5%)



Intuition: Omitted Variable Bias (OVB) (with *p*=1)

$$\sqrt{n}(\hat{\alpha}-\alpha)=\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{d_i\,\varepsilon_i}{\sqrt{n}}+\sqrt{n}\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{x_i^2}{n}\beta\gamma:=(A)+(B).$$

(drop *x*)

with $\beta = O(\frac{1}{\sqrt{n}})$

$$\sqrt{n}\beta\gamma \approx \sqrt{n}\,O(\frac{1}{\sqrt{n}})\,\gamma \nrightarrow 0 \qquad \text{if } \gamma \neq 0 \quad (PS)$$

but

$$\sqrt{n}\beta\,\gamma \approx \sqrt{n}\,O(\frac{1}{\sqrt{n}})\,O(\frac{1}{\sqrt{n}}) \to 0, \quad \text{if } \gamma = O(\frac{1}{\sqrt{n}}) \quad (DPS)$$

Similar with IV

DGP

$$
\begin{aligned}
y_i &= \alpha d_i + \beta x_i + \varepsilon_i, \\
d_i &= \gamma z_i + e_i, \\
z_i &= \varphi x_i + \omega_i,
\end{aligned}
$$

with $E(\varepsilon_i\, e_i) \neq 0, E(\varepsilon_i\, \omega_i) = E(\varepsilon_i\, x_i) = E(\,\omega_i x_i) = 0$

- DPS selects x if it´s a good predictor for y, d, z
- Then, apply 2SLS with chosen $x´s$
- Both for OLS and IV $\widehat{\sigma}_n^{-1}\sqrt{\widehat{\alpha} - \alpha} \rightsquigarrow N(0,1)$.

Selection in nonlinear models

DGP

$$\begin{aligned} y &= \alpha_0 d + g_0(\mathbf{x}) + \varepsilon, \\ d &= m_0(\mathbf{x}) + \nu, \end{aligned}$$

(partially linear)

Robinson (1988)

with $E(\varepsilon / d, x) = E(v / x) = 0$

$$\widehat{\alpha}_{PS} = \left(\sum_{i=1}^{n} \frac{d_i^2}{n} \right)^{-1} \sum_{i=1}^{n} \frac{d_i (y_i - \widehat{g}_0(x))}{\sqrt{n}},$$

$$\sqrt{n}(\widehat{\alpha}_{PS} - \alpha_0) = \left(\sum_{i=1}^{n} \frac{d_i^2}{n} \right)^{-1} \sum_{i=1}^{n} \frac{d_i \varepsilon_i}{\sqrt{n}} + \left(\sum_{i=1}^{n} \frac{d_i^2}{n} \right)^{-1} \sum_{i=1}^{n} \frac{d_i (g_0(x) - \widehat{g}_0(x))}{\sqrt{n}} := (A) + (B).$$

Nonparametric convergence rate of $\hat{g}_0(x)$ is $n^{-\varphi}, 0.25 < \varphi < 0.5$

$$(B) \approx \sqrt{n} n^{-\varphi} \to \infty,$$

$$\widehat{\alpha}_{PDS} = \left(\sum_{i=1}^{n} \frac{\widehat{v}_i^2}{n} \right)^{-1} \sum_{i=1}^{n} \frac{\widehat{\varepsilon}_i \widehat{v}_i}{\sqrt{n}},$$

$$\begin{aligned} \sqrt{n}(\widehat{\alpha}_{PDS} - \alpha_0) &= \left(\sum_{i=1}^{n} \frac{v_i^2}{n} \right)^{-1} \sum_{i=1}^{n} \frac{\varepsilon_i \nu_i}{\sqrt{n}} + \left(\sum_{i=1}^{n} \frac{v_i^2}{n} \right)^{-1} \sum_{i=1}^{n} \frac{(l_0(\mathbf{x}) - \widehat{l}_0(x))(m_0(\mathbf{x}) - \widehat{m}_0(\mathbf{x}))}{\sqrt{n}} + r(\mathbf{x}), \\ &:= (A) + (B) + (C) \end{aligned} \tag{17}$$

$$(B) \approx \sqrt{n} n^{-(\varphi_m + \varphi_l)} \to 0.$$

$(C) = o(1)$ *if sample split in* 2 *parts*: (i)*one to estimate* $l_0(x)$ & $m_0(x)$ and (ii)*another to regress* $\hat{\varepsilon}$ on $\hat{v}$

All are cases of Neyman Orthogonality Condition

$$\delta_\eta \mathbb{E}(\Psi(W, \theta_0, \eta)_{\eta=\eta_0} = 0$$

with $\theta = \alpha$ & $\eta = (l_0, m_0)$

DGP

$$y = h_0(d, \mathbf{x}) + \varepsilon,$$
$$d = m_0(\mathbf{x}) + \nu,$$

(non linear)

Chernozhukov et al (2017)

$$with\ d = \{0,1\}$$

$$\theta_0 = \mathbb{E}[h_0(1,\mathbf{x}) - h_o(0,\mathbf{x})] = \frac{d\,y}{\Pr(d=1 \mid \mathbf{x})} - \frac{(1-d)\,y}{1-\Pr(d=1 \mid \mathbf{x})},$$

$$\Psi(W,\theta,\eta) = h(1,x) - h(0,\mathbf{x}) - \frac{d\,(y-h(1,\mathbf{x}))}{m(\mathbf{x})} - \frac{(1-d)\,(y-h(0,\mathbf{x}))}{1-m(\mathbf{x})} - \theta,$$

$$with\ \eta = (h(1,x), h(0,x), m(x))$$

- Use a K-validation approach with $\widehat{\theta}_{0K} = K^{-1}\Sigma_{k=1}^{K}\theta_0(I_k, I_k^c)$

$$\sigma^{-1}\sqrt{n}(\widehat{\theta}_{0K} - \theta_0) \rightsquigarrow N(0,1),$$

A model of interactions: Modified Covariates Method (MCM)

Tian et al (2014), Chen et al. (2017)

$$y_i = \beta' \mathbf{x}_i + d_i \, \delta' \mathbf{x}_i + \varepsilon_i,$$

$$with\ d = \{0,1\} \Rightarrow T = 2d - 1 \Rightarrow \frac{T}{2} = \{-0.5, 0.5\}$$

$$y_i = \beta' \mathbf{x}_i + \frac{T_i}{2} \delta' \mathbf{x}_i + \varepsilon_i.$$

$$Cov(x_{ij}, T_i x_{ik}) = Cov(x_{ij}, x_{ik}) E(T_i) = 0$$

Use Lasso in: $y_i = \frac{T_i}{2} \delta' \mathbf{x}_i + \varepsilon_i,$

Empirical application (AK (1991)

$$
\begin{aligned}
w_i &= \alpha s_i + \beta' \mathbf{x}_i + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid \mathbf{x}, \mathbf{z}) = 0. \\
s_i &= \gamma' \mathbf{z}_i + \phi' \mathbf{x}_i + \nu_i, \ \mathbb{E}(\nu_i \mid \mathbf{x}, \mathbf{z}) = 0,
\end{aligned}
$$

z= **quarters of birth**, year of birth, state dummies, double & triple interactions (#1530)

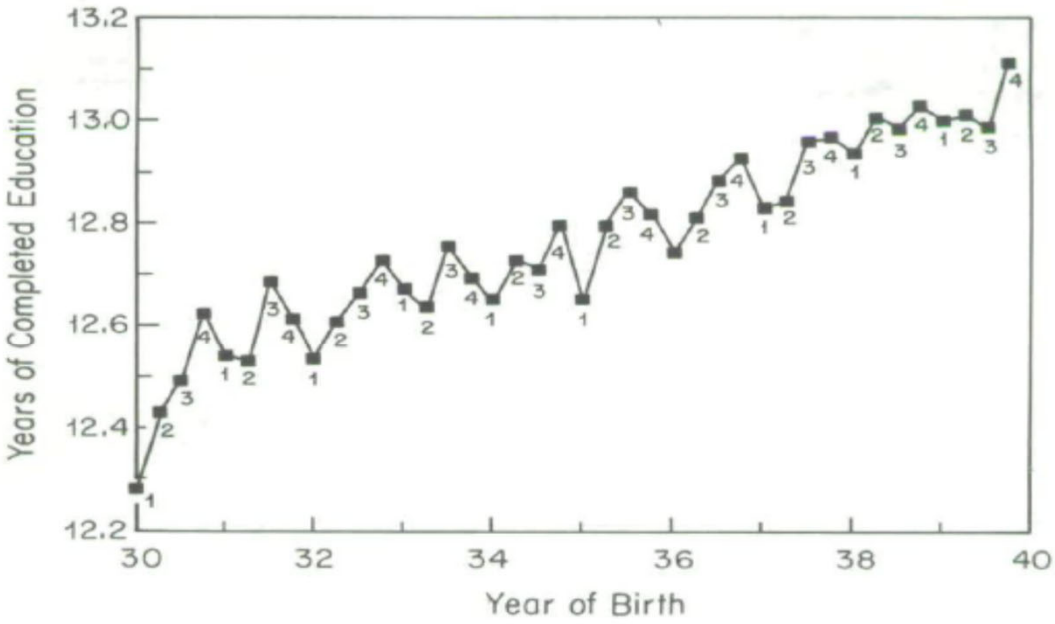


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

Returns to education (original IVs + Lasso)

Cuadro 1. Estimación del Rendimiento de la Educación en AK (1991)

nº IVs	MCB	FULL
3	0.106 (0.019)	0.109 (0.021)
180	0.096 (0.010)	0. 103 (0.014)
1530	0.069 (0.005)	0.108 (0.042)
1 (*)	0.087 (0.031)	—
12(**)	0.088 (0.014)	0.089 (0.014)

Nota: Estimadores MCB y FULL del parámetro α en el modelo de regresión (21).

Desviación típica entre paréntesis. (*) Lasso plug–in, (**) Lasso CV

Stata commands: poivregress and lasso linear

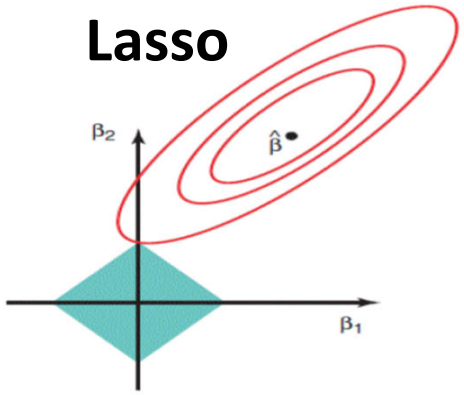
APPENDIX

ML for dummies

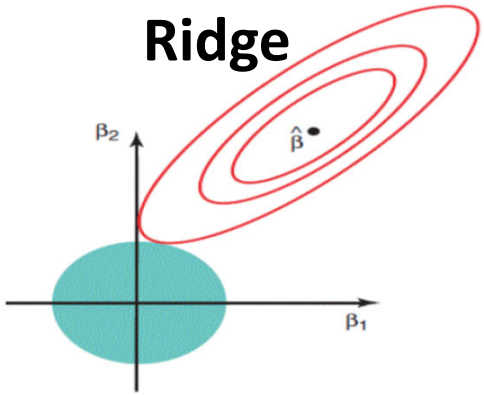
Ridge & Lasso

$$\min_b E\{Y_i - X_i'b\}^2 + \lambda||b||_q^q, \qquad ||b||_q = (\textstyle\sum_k |b_k|^q)^{1/q}.$$

Lasso



Ridge



Lasso **Adaptive Lasso**

$$\min_b \left[\sum_{i=1}^{n} \frac{(y_i - \sum_{j=1}^{p} b_j\, x_{ij})^2}{n} + \frac{\lambda}{n} \sum_{j=1}^{p} |b_j| \right], \quad \min_b \left[\sum_{i=1}^{n} \frac{(y_i - \sum_{j=1}^{p} b_j\, x_{ij})^2}{n} + \frac{\lambda}{n} \sum_{j=1}^{p} \varpi_j |b_j| \right].$$

Root Lasso

$$\min_b \left[\sqrt{\sum_{i=1}^{n} \frac{(y_i - \sum_{j=1}^{p} b_j x_{ij})^2}{n}} + \frac{\lambda}{n} \sum_{j=1}^{p} |b_j| \right],$$

Tuning parameters in Lasso

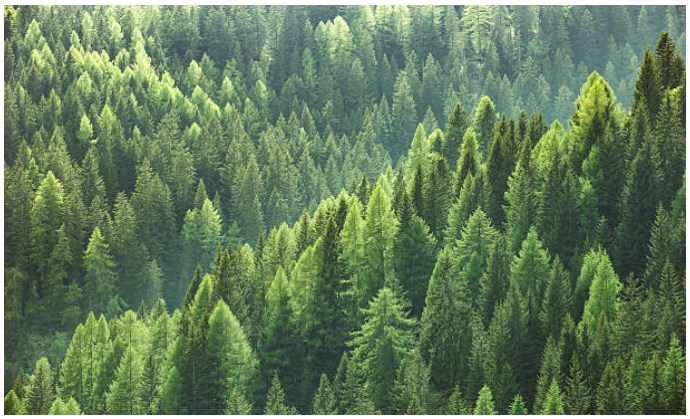
- $\Pr(\lambda \geq c \max_{1<j<p} \left|\frac{1}{n}\left(\frac{1}{\varpi_i}\right)\sum_{i=1}^{n} q_i\,(y_i, x_i, \theta_0)\right|) \longrightarrow 1$

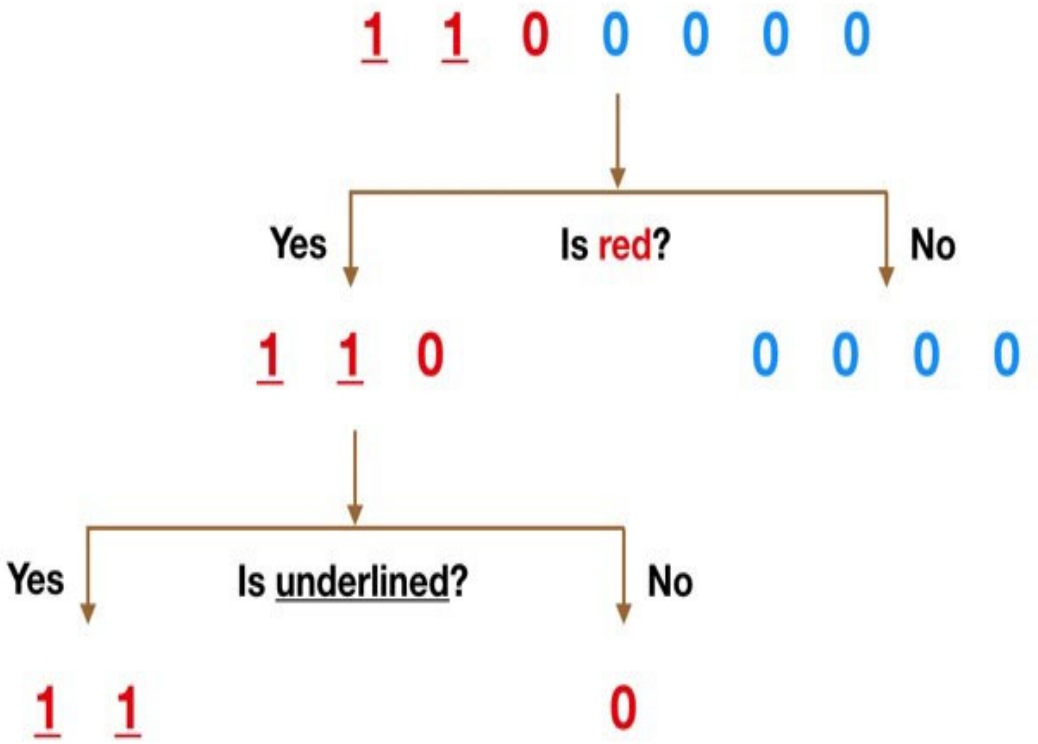
  where $q = \partial Q / \partial \theta$ and $c > 1$

$$\Pr(\lambda \geq c \max_{1<j<p} \left|\frac{1}{n}\left(\frac{1}{\varpi_i}\right)\sum_{i=1}^{n} q_i\,(y_i, x_i, \theta_0) \leq \Phi^{-1}(1 - \frac{\gamma}{2p})\right| \geq 1 - \gamma + o(1)$$

$$\Longrightarrow \lambda = \frac{c}{\sqrt{n}}\,\Phi^{-1}\left(1 - \frac{\gamma}{2p}\right) \quad \& \; \varpi_i = \sqrt{\frac{1}{n}\sum_{i=1}^{n} q_i^2\,(y_i, x_i, \theta_0)}$$

Trees and Forests





Thanks for your attention