Estimating intergenerational income mobility on sub-optimal data: a machine learning approach

ZEW 22-23 Mar. 2021

SOCIAL MOBILITY AND ECONOMIC PERFORMANCE

Paolo Brunori

University of Florence & University of Bari

This presentation

- A paper coauthored with:
  Francesco Bloise and Patrizio Piraino

- Should machine learning replace traditional methods when estimating intergenerational earnings elasticity (IGE)?

- Yes*

* unless longitudinal/administrative data available.

Two-sample two-stage least squares estimator (Björklund and Jäntti,1997)

$$y_i^c = \alpha + \beta y_i^p + \epsilon_i$$

- Impossibility to observe the fathers-sons link in the data:
  1. main sample (children);
  2. auxiliary sample (pseudo-fathers).

Two-sample two-stage least squares estimator (Björklund and Jäntti,1997)

- First stage: $y_{it}^{ps} = \phi z_i^{ps} + \theta_{it}$
- $\hat{y}_i^p = \hat{\phi} z_i^p$
- Second stage: $y_i^c = \beta \hat{y}_i^p + \omega_i$

Sources of bias

- Assuming to correctly measure y_i^c (Haider and Solon, 2006; Nybom and Stuhler, 2016) two additional biases (Solon, 1992; Björklund and Jäntti,1997):

  $\downarrow$ due to incorrect prediction of y_i^p;

  $\uparrow$ due to endogeneity of z.

  $\rightarrow$ the higher fist stage R^2 the higher the risk of a severe upward bias (Jerrim et al., 2016).

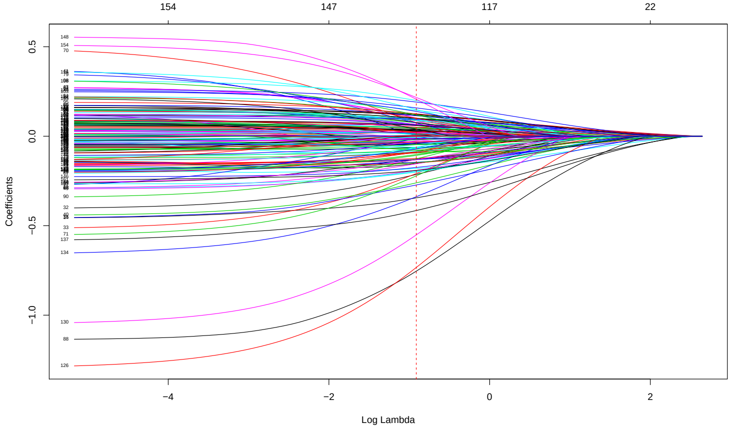
Data

- Panel Study of Income Dynamics (PSID);
- sons: adult individuals in 2011 for which we observe parental income for at least 5 waves between 1968 and 1992;
- "real" $IGE = 0.492$;
- pseudo-fathers: adults in 1982;
- z: education, occupation, industry, race (+ pairwise interactions) $\rightarrow$ 257 specifications.

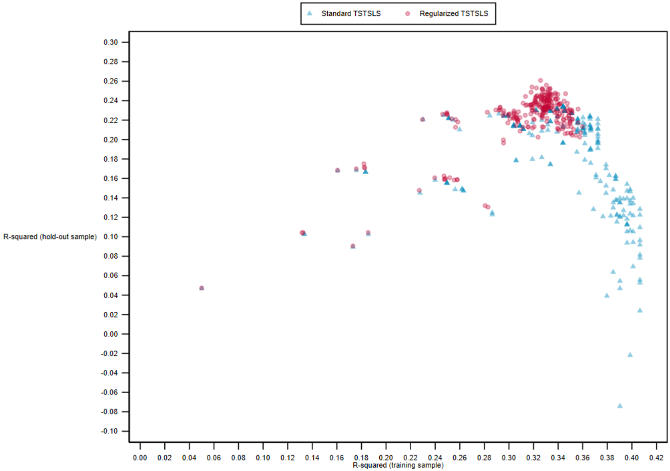
What we show

- $\hat{\beta}$ does not monotonically increase with R^2;
- selecting the model with ML reduces the upward bias;
- without incurring in substantial downward bias.

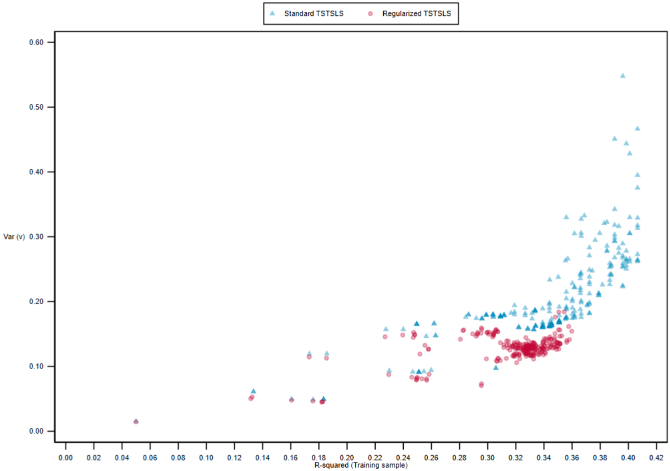
Relaxed elastic-net (Meinshausen, 2007; Hastie et al., 2017)



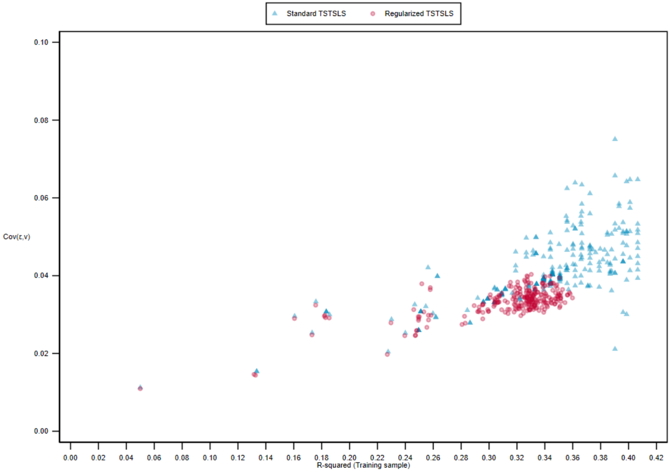
R^2 in-sample and out-of-sample



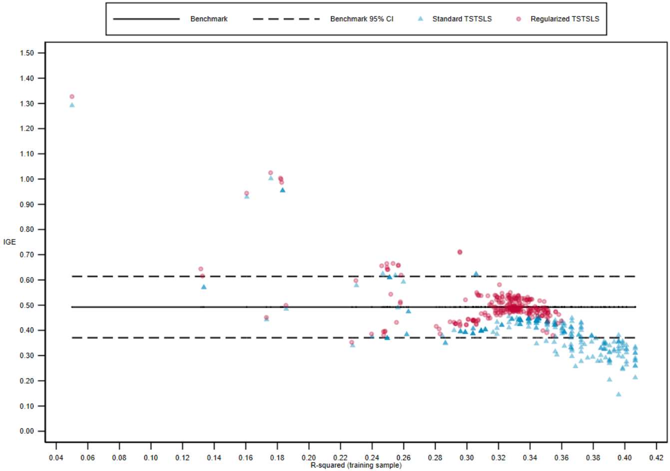
Downward bias (incorrect prediction of fathers' income)



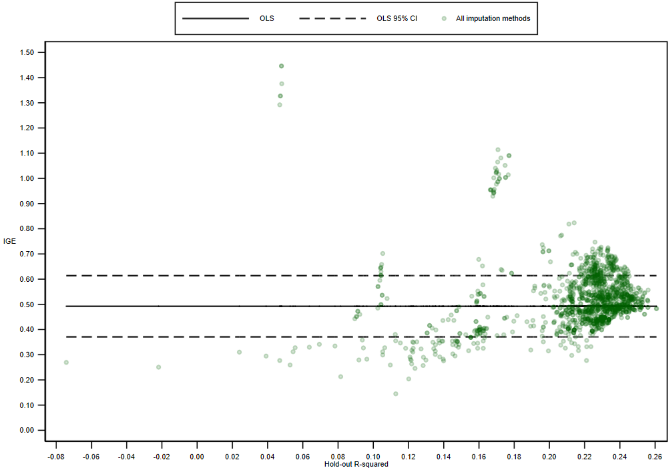
Upward bias (endogeneity)



$\hat{\beta}$: OLS vs. relaxed elastic net



Additional material: $\hat{\beta}$ alternative ML methods



Algorithms used: LASSO, ridge regression, elastic net, boosted regression, random forests.

Additional material: decomposition of the biases

$$\hat{y}_i^p = \gamma y_i^p + v_i$$

$$\text{plim } \beta_{TSTSLS} = \frac{cov(y_i^c, \hat{y}_i^p)}{var(\hat{y}_i^p)} =$$

$$= \frac{\gamma cov(y_i^c, y_i^p)}{\gamma^2 var(y_i^p) + var(v_i)} + \frac{cov(y_i^c, v_i)}{\gamma^2 var(y_i^p) + var(v_i)}$$

$$\text{plim } \beta_{TSTSLS} = \theta\beta + \underbrace{\frac{cov(\epsilon_i, v_i)}{\gamma^2 var(y_i^p) + var(v_i)}}_{\text{bias 2}}$$

$$\theta = \underbrace{\frac{\gamma var(y_i^p)}{\gamma^2 var(y_i^p) + var(v_i)}}_{\text{bias 1}}$$