Estimating intergenerational income mobility on two samples: sensitivity to model selection

Opportunities, Mobility and Well-Being
Polish Academy of Sciences, July 8-9 2019

Paolo Brunori

University of Florence & University of Bari

joint work with Francesco Bloise (University of Rome 3) and Patrizio Piraino (University of Cape Town)

Intergenerational elasticity of earnings

- the literature is developing in two directions:

  1. access to improved databases (Chetty and coauthors);

  2. access to some data for a larger number of countries (World Bank - LSMS).

- our contribution: a criterion to maximize comparability with suboptimal data (Brunori, Peragine, Serlenga, 2019).

Intergenerational elasticity of earnings

$$y_i^c = \beta_0 + \beta y_i^p + \epsilon_i$$

- y_i^c is the logarithm of the child's permanent income;
- y_i^p is the logarithm of the parent's permanent income;
- β is the intergenerational elasticity of income (IGE).

Two-Sample Two-Stage Least Squares (TSTSLS)

- Björklund and Jäntti (1997);
- *main* sample: information on adult income and their parents' socio-economic characteristics;
- *auxiliary* sample: earlier survey reporting pseudo-fathers' income and same socio-economic characteristics.

TSTSLS: first step

$$y_i^{ps} = \gamma z_i^{ps} + \theta_i \tag{1}$$

y_i^{ps} is the income of the pseudo-parents;

z are instrumental (imputer) variables;

γ is estimated by OLS.

TSTSLS: second step

$$y_i^c = \beta_0 + \beta \hat{y}_i^p + \omega_i$$

where $\hat{y}_i^p = \hat{\gamma} z_i^p$;

z_i^p are characteristics of the real fathers;

and $\hat{\beta}_{TSTSLS}$ is IGE.

TSTSLS: biases

1. endogeneity:
$$y_i^c = \beta_0 + \beta y_i^p + \gamma_2 z_i^p + \epsilon_i \tag{2}$$

2. first-stage incorrect prediction : $(R^2 < 1)$.

Sensitivity to model specification

Using Jerrim et al. (2016) notation:

$$\text{Plim}\beta_{TSTSLS} = \beta + \gamma_2 \left(1 - R^2\right)$$

- the higher R^2, the lower the bias;
- the closer γ_2 to 0, the lower the bias.

Model selection

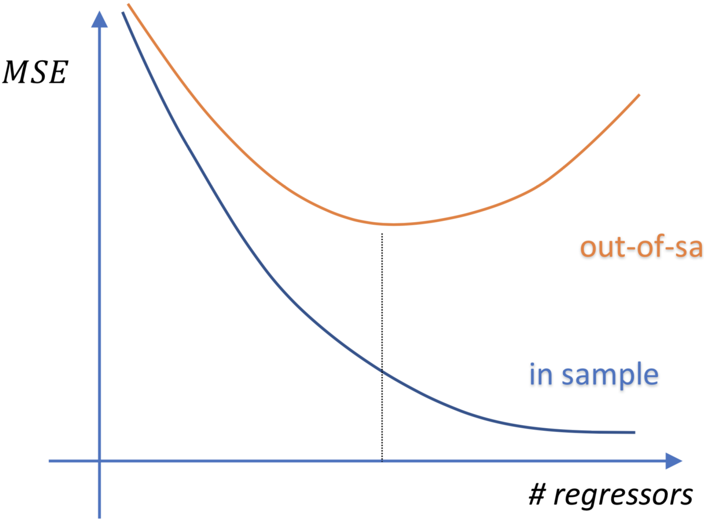
- larger R^2 improves our estimates?
- it can also increase γ_2;
- R^2 monotonically increase with # of regressors in sample
- but we are interested in predicting y of unseen fathers.
- the proper objective function si R^2 out-of-sample.

Model selection, cnt.

- maximizing ability to predict out-of-sample is what machine learning algorithms do.
- solving the bias-variance trade-off

$$MSE = E\left[(y_0 - \hat{f}(z_0))^2\right] = Var(\hat{f}(z_0)) + [Bias(\hat{f}(z_0))]^2 + var(\theta)$$

MSE out-of-sample

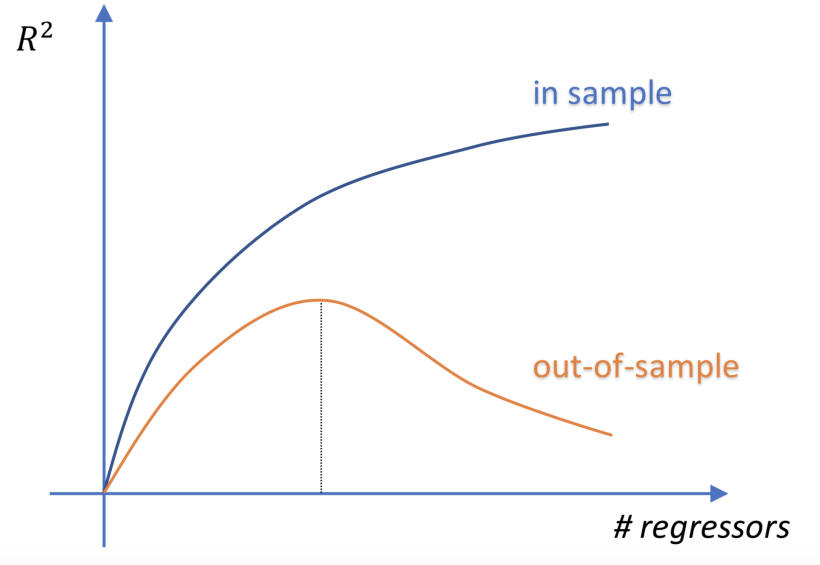


Model selection

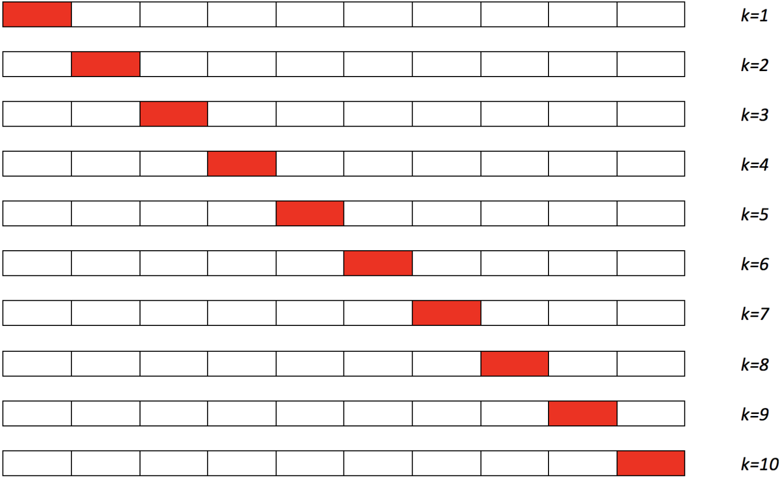
- to minimize MSE out-of-sample is equivalent to maximize R^2 out-of-sample

$$(1 - R^2) = n \frac{MSE}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

R^2 out-of-sample



k-fold cross validation



10-fold Cross Validation

Model selection

- Standard approach: specify a few linear and additive mode and discuss their credibility;
- two options:
  1. estimate MSE for all possible models (feasible in this case);
  2. regularization of linear models with interactions.

Regression regularization

- OLS search for the parameters that minimize MSE in sample;

- shrinking methods search for parameters that minimize MSE out-of-sample;

- general approach: penalize models with many parameters and models with large coefficients.

Ridge regression

Ridge regression shrinks regression coefficients by imposing a penalty on their size:

$$\hat{\beta}_{RIDGE} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3}$$

Ridge regression

Ridge regression shrinks regression coefficients by imposing a penalty on their size:

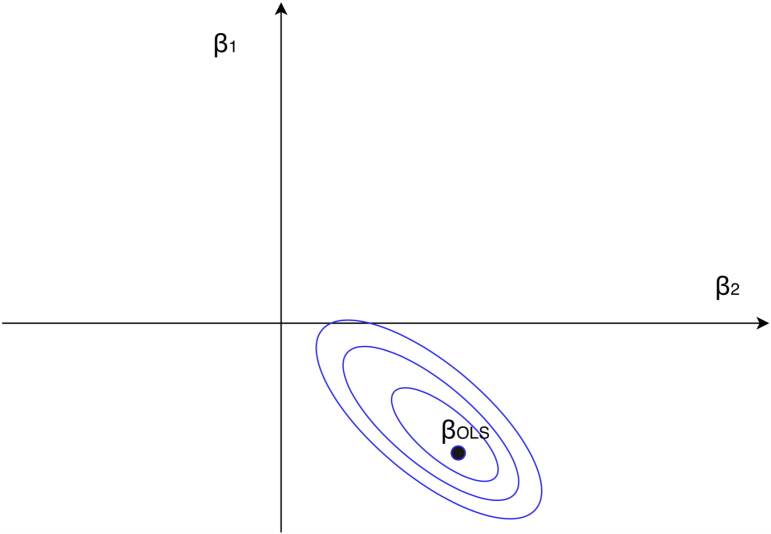
$$\hat{\beta}_{RIDGE} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{4}$$

This is equivalent to:

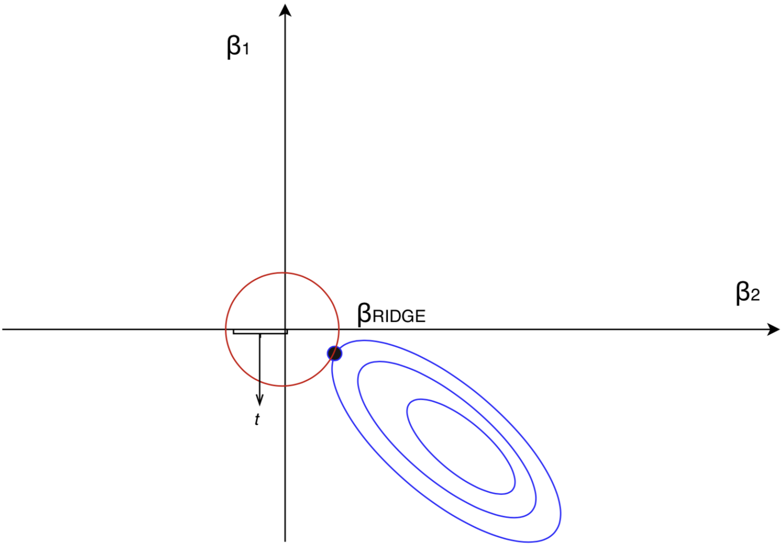
$$\hat{\beta}_{RIDGE} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t$$

OLS



Ridge regression



Regression regression

- contrary to other parsimony criteria (BIC, AIC) λ is not predetermined

- ridge regression is *tuned* searching for λ that produces lowest out-of-sample MSE by cross-validation

Least absolute shrinkage and selection operator (Lasso)

Lasso performs both variables selection and shrinkage by imposing a penalty on their absolute size:

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{5}$$

Lasso

Lasso shrinks regression coefficients by imposing a penalty on their absolute size:

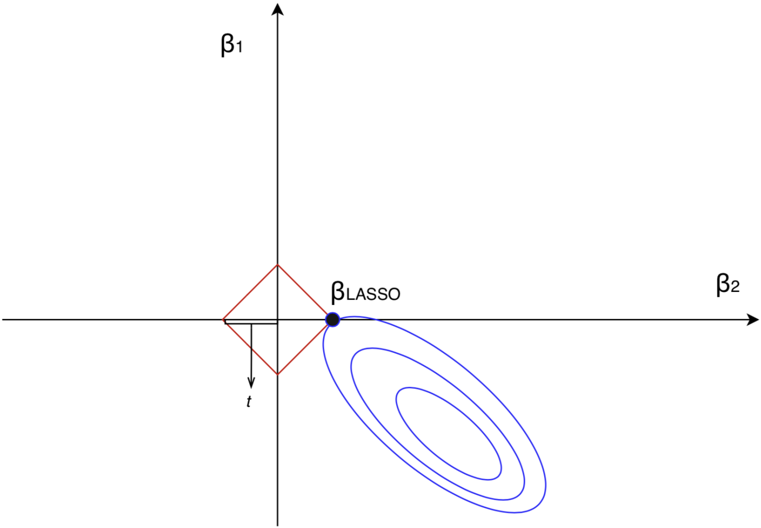
$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{6}$$

This is equivalent to:

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

Lasso



Lasso

- Lasso is also *tuned* searching for λ that produces lowest out-of-sample MSE by cross-validation;

- The non linearity of the constraint forces some coefficient to be exactly zero (a variables selection alogirthm);

- Zou and Hastie (2005) have proposed a to use a weighted average of the two methods: *elastic net*.

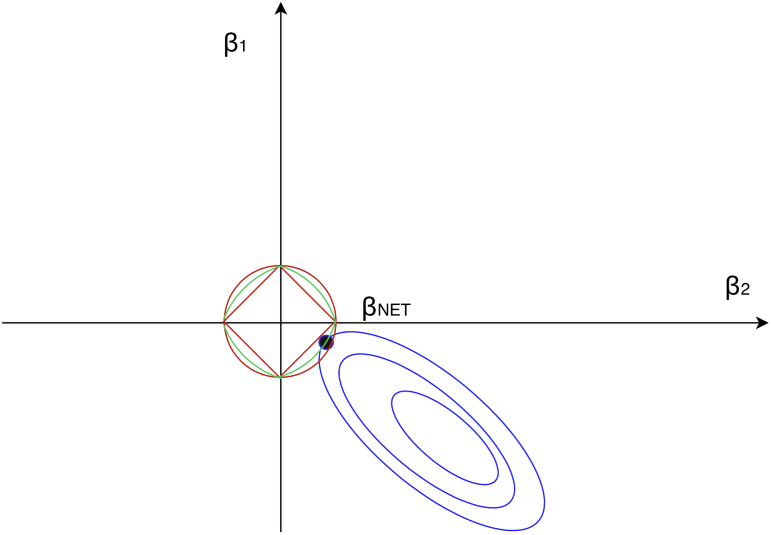
Elastic net

Elastic net is a weighted average of Lasso and ridge algorithm:

$$\hat{\beta}_{NET} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2 \right\} \tag{7}$$

$$\text{subject to} : (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t$$

Elastic net



Elastic net

- Tuning the elastic net implies searching for the couple α and λ that minimizes MSE:
- when $\alpha = 0$ we are back to ridge regression;
- when $\alpha = 1$ we are using a Lasso;
- when $\lambda = 0$ we are using standard OLS.

Data - US

- main sample: wave 2011 - Panel Survey of Income Dynamics (PSID);
- 1,061 sons, aged 30-60, with positive earnings and non-missing background information about their fathers;
- auxiliary sample of 1,860 pseudo-fathers aged 30-60 with positive earnings using the 1982 wave of the PSID.

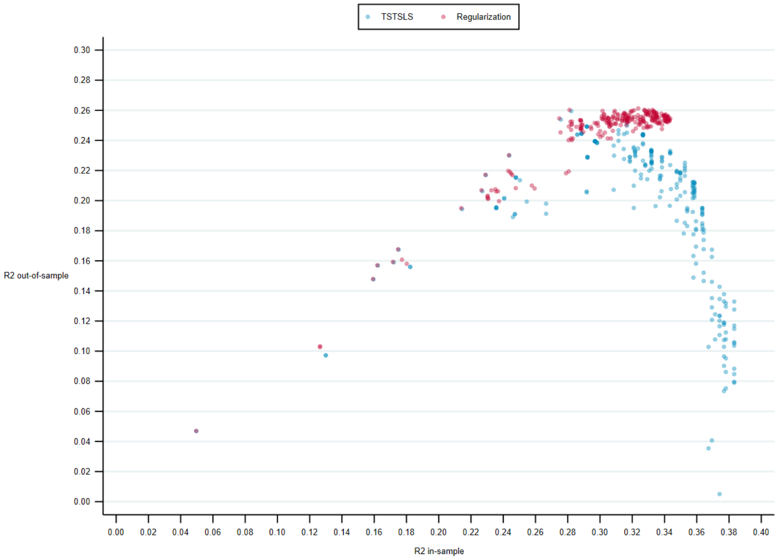
Data

- first-stage variables: education, occupation, industry, and race, plus all possible pairwise interactions (1,023 models);

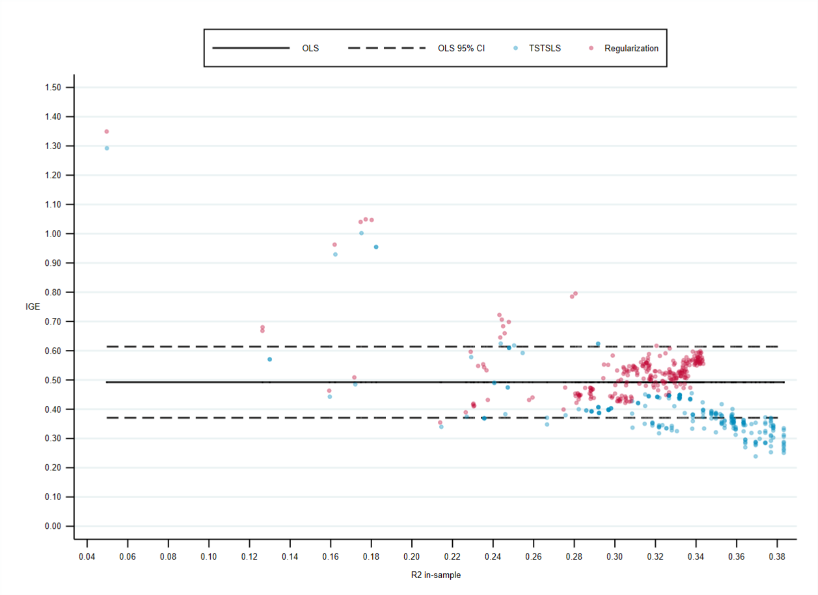
- update Björklund and Jäntti (1997);

- obtain benchmark longitudinal IGE.

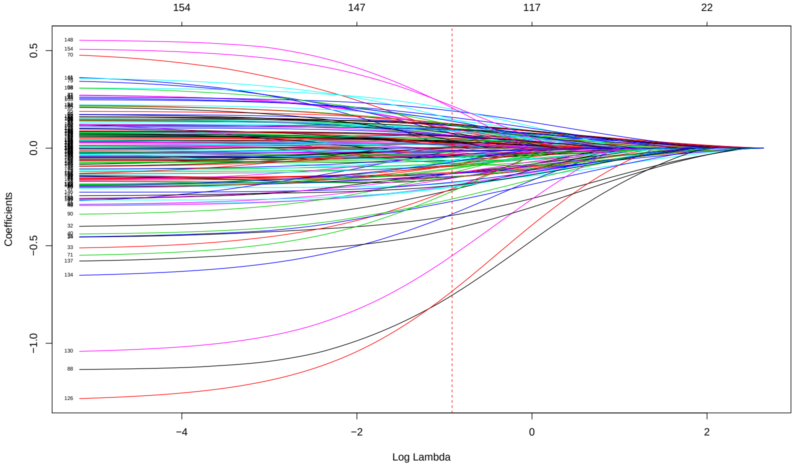
Model complexity and out-of-sample R^2



Model complexity and β_{TSTSLS}



Elastic net regularization



$\alpha = 0.010, \ \lambda = 0.402.$

β_{TSTSLS} sensitivity to model specification

	IGE	s.e.	First-Stage R^2 (out-of-sample)	First-Stage R^2 (in-sample)
Benchmark	0.492	(0.062)		
B&J, 1997	0.478	(0.073)	0.205	0.222
Best model	0.496	(0.078)	0.261	0.324
top 5 models	0.487	(0.074)	0.260	0.317
top 10 models	0.494	(0.080)	0.260	0.319
Sample size	1,061	1,061	1,860	1,860

Conclusions

- non-arbitrary selection criterion that produces non-trivial change in IGE;
- e.g. South Africa $0.62 \to 0.69$;
- open question: under what condition regularization does not exacerbate upward bias due to endogeneity?