Working Paper Series

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

Francesco Bloise
Paolo Brunori
Patrizio Piraino

ECINEQ WP 2020 - 526

ECINEQ 2020 - 526
March 2020

www.ecineq.org

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

Francesco Bloise
University of Rome Tre

Paolo Brunori
University of Florence and University of Bari

Patrizio Piraino
University of Notre Dame

Abstract

Much of the global evidence on intergenerational income mobility is based on sub-optimal data. In particular, two-stage techniques are widely used to impute parental incomes for analyses of developing countries and for estimating long-run trends across multiple generations and historical periods. We propose a machine learning method that may improve the reliability and comparability of such estimates. Our approach minimizes the out-of-sample prediction error in the parental income imputation, which provides an objective criterion for choosing across different specifications of the first-stage equation. We apply the method to data from the United States and South Africa to show that under common conditions it can limit the bias generally associated to mobility estimates based on imputed parental income.

Keywords: Intergenerational elasticity; income; mobility; elastic net; regularization; PSID, South Africa.

JEL Classification: J62; D63; C18.

1. Introduction

A large empirical literature in economics focuses on the estimation of the degree of income persistence across generations. Studies in this literature typically estimate simple regression models that deliver an estimate of the statistical association between the income of parents and that of their adult offspring. Although no causal interpretation is possible, these correlations are generally used as informative statistics for the level of social mobility within a country—see Corak (2013) and Emran and Shilpi (2019) for reviews.

Despite the clear relevance of intergenerational economic mobility to equity, efficiency and public policy, economists have only recently renewed their interest in the issue. During the last three decades, increased access to data has enabled multiple years of observations of the economic status of successive generations in a number of countries. In addition, new methodological tools have allowed a clearer understanding of the key measurement issues in assessing the intergenerational transmission of economic status. In high-income countries, and in an increasing number of low and middle-income countries, the new empirical analyses have allowed comparisons of the extent of social mobility across nations with different economic systems and values (Solon, 2002; Björklund and Jäntti, 2009) as well as over time and space for a subset of countries (Aaronson and Mazumder, 2008; Olivetti and Paserman, 2015). These comparisons have shown significant variation in the degree of intergenerational income inequality, thereby paving the way for the investigation of the institutional and policy features that can help explain the observed patterns (Blanden, 2013; Chetty et al., 2014).

At the same time, it is noticeable that the global evidence on intergenerational income mobility is often based on low-quality data. These are instances where the available observations do not permit to establish a direct parent-child link with adequate income information. This limitation is of particular relevance for developing countries and for historical analyses of mobility in societies at various stages of economic development. The widespread use of sub-optimal data affects the credibility of comparative analyses, to the extent that differences in observed levels of mobility may be driven by varying data conditions (Emran and Shilpi, 2019). The contribution of our paper is to propose an estimation approach that can improve the reliability and comparability of intergenerational mobility estimates based on sub-optimal data.

Specifically, we propose a modification of the current workhorse estimator used in the literature for measuring mobility when intergenerationally-linked income information is not available. This is the Two-Sample Two-Stage Least Squares (TSTSLS) estimator originally

pioneered by Björklund and Jäntti (1997) and used since then in numerous empirical studies (e.g. Aaronson and Mazumder, 2008; Gong et al., 2012; Olivetti and Paserman, 2015; Piraino, 2015). This estimator uses retrospective information on socioeconomic background along with a sample of 'pseudo' parents to impute parental incomes. Since background information of this type is more likely to be available in survey datasets (or historical censuses) compared to parental income, the TSTSLS methodology allowed the estimation of intergenerational income mobility for a significantly larger number of countries and historical periods, with a major impact in the coverage of developing nations (Narayan et al, 2018; Brunori et al., 2020).

We advance an approach that improves the imputation of parental income in the TSTSLS and that provides an objective criterion for choosing across different specifications of the prediction equation. By taking advantage of machine learning techniques, we minimize the out-of-sample prediction error. Using a criterion that is applicable to different data conditions can increase the comparability across studies, as mobility estimates become less sensitive to arbitrary specification choices. Since it is not possible to know *a priori* which model best predicts parental income in different contexts, we suggest a data-driven routine for model selection in the first stage of the TSTSLS. Researchers working on (potentially) very different datasets can utilize the same algorithm, which exploits the information embedded in all available predictors of parental income. Among the algorithms currently available to minimize the out-of-sample prediction error, we opt for a shrinkage method (Zou and Hastie, 2005). The method avoids overfitting by shrinking the standard linear regression coefficients. Unlike other algorithms, regularized regression improves the precision of the estimates without limiting our ability to easily interpret the output.

We show the usefulness of our methodological approach by testing its performance on longitudinal income survey data from the United States (PSID). The empirical analysis shows that our method reduces the distance between the TSTSLS estimate and the benchmark OLS estimate obtained from longitudinally-linked data on the *same* sample of individuals and their *real* parents. As noted in some recent studies (Olivetti and Paserman, 2015; Santavirta and Sthuler, 2019) and contrary to what is generally assumed in the earlier literature on intergenerational mobility (Corak, 2006), we confirm that the TSTLS estimator can produce both upward and downward biased estimates of the underlying true elasticity. This depends on the relative magnitude of the downward bias induced by measurement error in imputed incomes and the upward bias due to the residual association (i.e. uncorrelated with parental income) between first-stage predictors and child's income. By virtue of focusing on the maximum predictive power (out-of-sample) of the first stage, our approach limits both measurement error

and the predictors' informational content over and above parental income. By constraining both sources of bias, which move in opposite directions, the algorithm limits the risk of TSTSLS delivering an estimate overly affected in either direction.

We test the applicability of our method to sub-optimal data conditions by replicating part of the analysis on survey data from South Africa. While we do not have a benchmark longitudinal estimate on this sample, the estimator produces analogous results for the subset of estimates we can reproduce. Taken together, our findings on the United States and South Africa are of high relevance for the vast majority of countries (and of the world's population) where long-span income information, from either administrative or survey panel data, is not available. More generally, we suggest that machine learning approaches, such as the one advanced in this paper, should become part of the standard set of empirical tools for analyses of intergenerational income mobility relying on imperfect data.

The rest of the paper proceeds as follows. Section 2 revisits the standard TSTSLS estimator and clarifies its sources of bias. Section 3 presents our machine learning method. Section 4 shows the empirical results, while Section 5 concludes.

2. Two-sample two-stage least squares (TSTSLS) estimator

The standard empirical specification for estimating intergenerational income mobility is given by the following equation:

$$y_i^c = \alpha + \beta y_i^p + \epsilon_i \quad (1)$$

where y_i^c is the logarithm of the child's permanent individual income and y_i^p is the logarithm of the parent's permanent individual income. The coefficient estimate for β is generally named the 'intergenerational elasticity' (IGE) and forms the basis for comparisons across countries around the world.

Amongst the existing IGE estimates in the literature, a significant number (and virtually all of those for developing countries) are obtained through the TSTSLS methodology introduced by Björklund and Jäntti (1997). This estimation requires two samples. The *main* sample contains information on individual incomes and recall socioeconomic information about their parents. The *auxiliary* sample is typically derived from an earlier survey of the same

population where individuals (*pseudo-parents*) report their income as well as information similar to that recalled by respondents in the main sample.[1]

The estimation then proceeds in two steps. First, the auxiliary sample is used to estimate a Mincer equation:

$$y_{it}^{ps} = \varphi z_i^{ps} + \vartheta_{it} \tag{2}$$

where y_{it}^{ps} is the income of pseudo-parents in a given year, z_i^{ps} is a vector of time-invariant characteristics, and ϑ_{it} is the component of pseudo-parents' income that is not captured by the observed predictors. In the second step, the main sample is used to estimate the equation:

$$y_i^c = a + \beta \hat{y}_i^p + \omega_i \tag{3}$$

where y_i^c is the income of children. $\hat{y}_i^p = \hat{\varphi} z_i^p$ is the imputed income of *unseen* parents, and z_i^p are recall variables analogous to z_i^{ps}. Note that Equation (3) abstracts from measurement error in the child's permanent income. While left-hand side measurement error is a well-documented source of bias for the IGE (Haider and Solon, 2006; Nybom and Stuhler, 2016), our focus here is on the correct prediction of parental income.[2]

2.1 Sources of bias in TSTSLS estimates

Since intergenerational regression models do not aim to identify the causal effect of parental income on child income, the first-stage predictors need not satisfy any exclusion restriction. The sources of bias we discuss here refer to the difference between the TSTSLS estimate from Equation (3) and the elasticity estimated on Equation (1) under ideal data conditions (i.e. direct parent-child link and permanent incomes for both generations).

[1] A growing recent literature makes use of surnames or first names to impute parental socioeconomic status and estimate intergenerational mobility over the long-run in certain countries (e.g. Clark, 2014; Olivetti and Paserman, 2015). While these studies also use a TSTSLS (or related) estimator, our discussion here focuses on a scenario common to several contemporary developing countries, where survey data with recall information on parental background is available. The general idea of using machine learning to predict parent's income, however, extends to the set of studies using the informational content of (sur)names.

[2] In fact, Equations (1) to (3) may vary depending on data availability. Many of the existing IGEs in the literature, including most longitudinal OLS estimates, are based on imperfect measures of the child's permanent income.

Relative to the linked estimator on longitudinal data, the IGE obtained from the two-sample approach will suffer from two main sources of bias (Solon, 1992; Björklund and Jäntti, 1997; Jerrim et al., 2016):

(i) incorrect prediction of the income of unseen parents;

(ii) first-stage predictors entering the child's income equation over and above parental income.

Given the type of first-stage variables usually available to researchers (parental education, occupation, area of birth, etc..) it is common to treat TSTSLS estimates as *upper bound* values of the 'true' IGE. This is because the first-stage predictors are positively related to child income independently of parental income—i.e., bias *(ii)* is positive. Most studies providing TSTSLS estimates are less explicit about bias *(i)*, which may work in the opposite direction. The choice of the prediction model is generally motivated by data availability, and several IGE estimates based on different combinations of variables are presented as robustness checks. Thus, the sign of the overall bias in many of the existing TSTSLS estimates is a priori ambiguous.

In order to show how the approach we propose can limit the overall bias affecting the TSTSLS estimates, we derive a simple expression of the various components of the estimator. We begin by considering the linear projection of $\hat{y}_i^p$ on y_i^p:

$$\hat{y}_i^p = \gamma y_i^p + v_i \tag{4}$$

where v_i is the projection error.

Focusing on the right-hand side measurement error (i.e. assuming that child's permanent earnings are observable) we can use Equation (4) to express the probability limit of the TSTSLS estimator as follows:

$$plim\hat{\beta}_{TSTSLS} = \frac{cov(y_i^c,\hat{y}_i^p)}{var(\hat{y}_i^p)} = \frac{\gamma cov(y_i^c,y_i^p)}{\gamma^2 var(y_i^p)+var(v_i)} + \frac{cov(y_i^c,v_i)}{\gamma^2 var(y_i^p)+var(v_i)} \tag{5}$$

Which, using Equation (1), can be rewritten as

$$plim\hat{\beta}_{TSTSLS} = \theta\beta + \frac{cov(\epsilon_i,v_i)}{\gamma^2 var(y_i^p)+var(v_i)} \tag{6}$$

where $\theta = \frac{\gamma var(y_i^p)}{\gamma^2 var(y_i^p)+var(v_i)}$ represents bias (*i*), and the ratio $\frac{cov(\epsilon_i, v_i)}{\gamma^2 var(y_i^p)+var(v_i)}$ represents bias (*ii*). In general, bias (*i*) will be an attenuation bias as the denominator is greater than the numerator unless γ is extremely low.[3] Bias (*ii*) is typically assumed to be positive, which amounts to assuming that $cov(\epsilon_i, v_i) > 0$.

We show in the empirical analysis below how our method compares to the standard TSTSLS in terms of the size of both biases, which we are able to infer from our benchmark estimate on the longitudinal sample. Before turning to the empirical results, however, we first describe the machine learning approach used to minimize the out-of-sample prediction error in the parental income imputation (Equation 2).

3. Method

Our goal is to predict the earnings of *unseen* parents with the smallest possible squared error:

$$min\{\mathbb{E}[(y_0^p - \hat{y}_0^p)^2]\} = min\left\{\mathbb{E}\left[\left(y_o^p - f\left(z_0^{ps}\right)\right)^2\right]\right\} \quad (7)$$

where y_0^p is the income of the real parent of individual *0* (a person we do not observe) and $f\left(z_0^{ps}\right)$ is an unknown prediction function based on the vector z_0^{ps}. A well-known result in statistical learning is that, out-of-sample, the expected squared error of a prediction can be decomposed into three elements:

$$\mathbb{E}[(y_{ot}^{ps} - \hat{y}_{ot}^{ps})^2] = var\left(\hat{f}\left(z_0^{ps}\right)\right) + (bias)^2 + var(\vartheta_{0t}) \quad (8)$$

where $var\left(\hat{f}\left(z_0^{ps}\right)\right) = \mathbb{E}\left[\hat{f}(z_0^{ps})^2\right] - \mathbb{E}\left[\hat{f}\left(z_0^{ps}\right)\right]^2$ is the *variance* of the model; that is the error caused by the sensitivity of the model to random noise in the observed sample. The term $bias = \mathbb{E}\left[\hat{f}\left(z_0^{ps}\right)\right] - \mathbb{E}\left[f\left(z_0^{ps}\right)\right]$ is the *bias* of the model, which quantifies the error that is introduced by approximating an unknown data generating process by a simpler model (for

[3] The term describing bias *(i)* above is similar to the first term in Equation (2) in Olivetti and Paserman (2015), who derive the relationship between the TSTSLS (*pseudo-panel*) estimator and the longitudinal OLS (*linked*) estimator in the context of name-based imputations of parental economic status.

example by assuming additivity of the predictors' effect or excluding interaction effects). Finally, $var(\vartheta_{0t})$ is variation unrelated with covariates and is therefore an irreducible term of the out-of-sample prediction error. When trying to minimize Equation (7) on a limited number of observations, we face a trade-off. Very complex models will tend to have low *bias* and large *variance*. On the other hand, overly simple models are characterized by high *bias* and low *variance*. We handle such variance-bias trade off departing from the classical least square regression analysis and estimating the first-stage regression using the elastic-net shrinkage operator introduced by Zou and Hastie (2005). An elastic-net obtains the regression coefficients by minimizing:

$$\sum_{i=1}^{n}\left(y_i - b_o - b_1 X_{1,i} - b_2 X_{2,i} \ldots - b_k X_{k,i}\right)^2 + \lambda\left(\alpha \sum_{j=1}^{k} |\, b_j| + (1-\alpha) \sum_{j=1}^{k} {b_j}^2\right) \qquad (9)$$

The regularization term $\left(\alpha \sum_{j=1}^{k} |\, b_j| + (1-\alpha) \sum_{j=1}^{k} {b_j}^2\right)$ shrinks the coefficient estimates towards zero, in order to avoid the risk of overfitting. $\lambda \geq 0$ is a parameter that controls the importance of the regularization term. Elastic-net is a linear combination of two standard operators in machine learning: LASSO (least absolute shrinkage and selection operator) and ridge regression. When $\alpha = 0$, the elastic-net algorithm is equivalent to the ridge regression. When $\alpha = 1$, it is equivalent to the LASSO. Provided that $\lambda > 0$ and $\alpha > 0$, some coefficients will be set exactly to zero and others will be shrunk.

Using elastic-net, we obtain different sets of b_s depending on the value of λ and α. In statistical learning terminology, this implies that the algorithm needs to be tuned so as to obtain a more precise model specification. Among all possible specifications, we aim at tuning λ and α so that Equation (7) is minimized. A standard method to tune elastic nets is *k*-folds cross-validation. Cross-validation provides a direct estimate of the out-of-sample prediction error under very weak assumptions (Arlot and Celisse, 2010). A reader familiar to machine learning may wonder why we opt for the elastic-net, an algorithm that can be outperformed by others in term of predicting performance (James et al., 2013). The reason is that elastic-nets, contrary to more complex algorithms, are rather efficient and very easy to interpret, since they regularize an ordinary least squares regression.

4 Empirical analysis

We first provide an empirical application of our method using longitudinal survey data from the United States. This allows us to benchmark the performance of the estimator in a scenario where we can obtain the IGE through both a standard OLS on a single longitudinal sample and through the TSTSLS on two separate samples. We then replicate part of the analysis on South African data, which provides a case study for typical sub-optimal data conditions in the developing country literature.

4.1 Standard and regularized TSTSLS vs. benchmark longitudinal OLS

For the sake of simplicity, and consistent with a large section of the literature, we restrict our analysis to males only. For the United States, we use the 2011 wave of the Panel Survey of Income Dynamics (PSID) to obtain the *main sample* of sons aged 30-60, with positive earnings and non-missing background information about their fathers.[4] In the longitudinal OLS specification, the earnings of real fathers are averaged over all yearly observations available. We include only sons whose real fathers have at least five years of positive earnings (and were 30 to 60 years old) between 1968 and 1992. The final *main* sample consists of 1,061 observations.

We then obtain an *auxiliary* sample of 1,860 pseudo-fathers aged 30-60 using the 1982 wave of the PSID. In both the *main* and *auxiliary* samples, we use yearly gross employment income, constructed as the sum of wages, salary bonuses, overtime income, labor income from business, commission income, income from professional practice or trade and labor part of income from farming or market gardening.

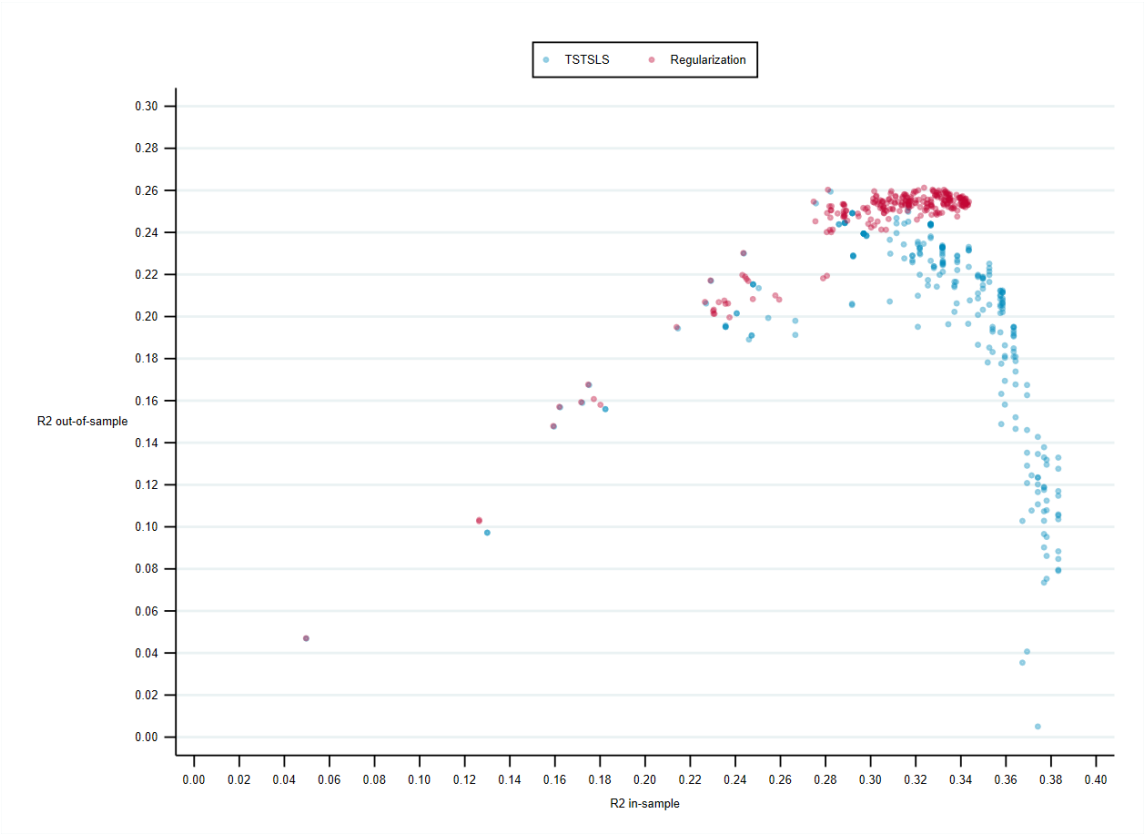
When estimating β_{TSTSLS}, it is common practice in the literature to use different additive combinations of the available first-stage predictors and report the resulting coefficients. Instead, our approach lets the elastic-net find the specification that minimizes the out-of-sample prediction error for each potential set of regressors. In our sample, the first-stage variables are dummies for education (8), occupation (9), industry (9), and race (3), plus all possible pairwise interactions. The regularization of the first-stage model is thus performed on

[4] As mentioned above, the imperfect measurement of child's income can introduce a bias in both the OLS and TSTSLS estimates. We select a sample of children who are on average 44.8 years old, which is line with the range suggested in the literature to minimize the left-hand side bias.

1,023 different models.[5] Amongst models with an equal number of regressors, we select the one with the highest R^2 (in-sample).[6] This results in 257 models of varying complexity (i.e. number of regressors) for which we estimate the in- and out-of-sample R^2 for both the regularized and standard TSTSLS. Figure 1 shows the relationship between the in-sample (x-axis) and out-of-sample (y-axis) R^2 for the estimated models.

Figure 1:

Increasing complexity of the first-stage model and predictive performance of standard TSTSLS vs. regularized regression



Source: PSID (1982-2011)
Notes: The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the out-of-sample R^2 estimated by 5-fold cross-validation for both the standard TSTSLS (blue) and regularized (red) models.

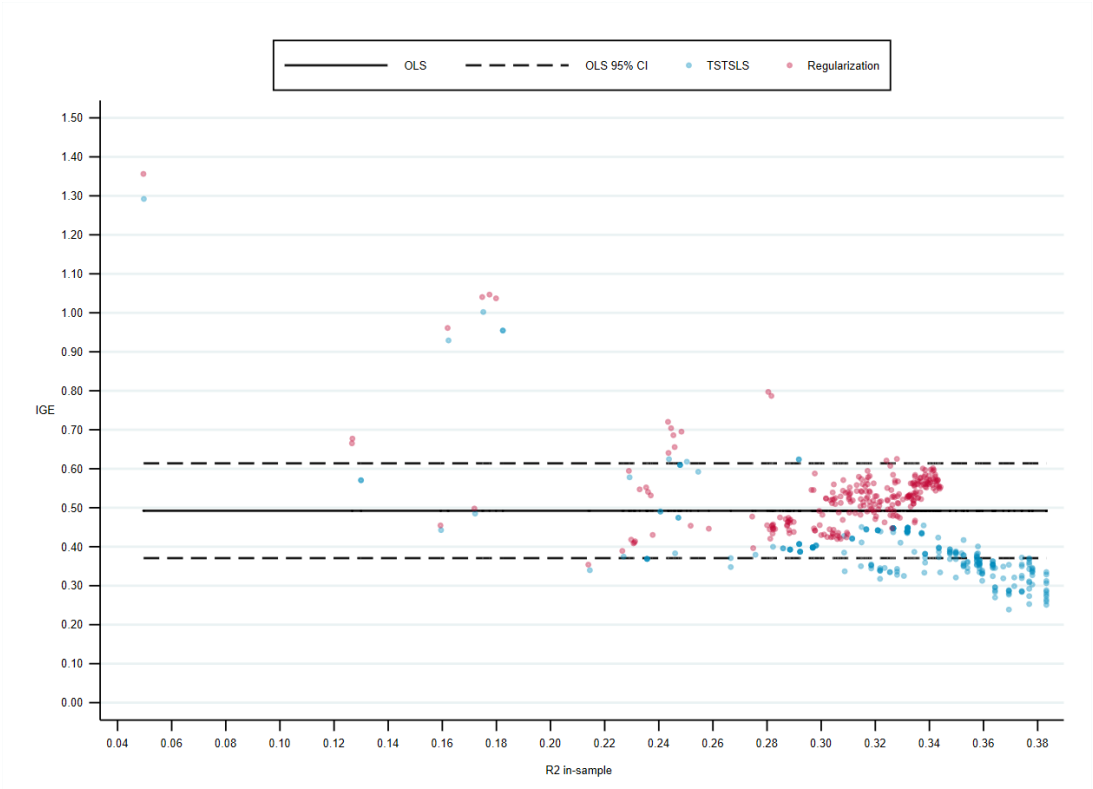
The first noticeable result from Figure 1 is that the predictive performance of the non-regularized regression (blue dots) shows the expected pattern: very parsimonious models (to the left of the graph) underfit the data while overly complex models (to the right) tend to overfit

[5] This is the sum of all k-combinations of the 10 available first-stage predictors (i.e. education, occupation, sector, race, education*occupation, education*race, education*sector, occupation*race, occupation*sector, race*sector).
[6] This "best subset regression" approach is a method to select the best performing model when, as in this case, the number of possible models is reasonably low. For a given number of controls (degrees of freedom), the in-sample prediction performance has a monotonic relationship with the out-of-sample performance. Therefore, it is sufficient to focus on models with the highest in-sample R^2.

the data, which reduces the ability to correctly predict out-of-sample. On the other hand, the regularized models (red dots), while performing worse in-sample, have significantly higher out-of-sample predictive power for more complex models as they are able to avoid overfitting. Our first result is thus to confirm that as models become more complex, regularization improves out-of-sample prediction.

Figure 2:

In-sample R^2 and estimated IGE for standard TSTSLS and regularized model (red)



Source: PSID (1982-2011)
Notes: The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the corresponding IGE estimate for both standard TSTSLS (blue) and regularized (red) models. The solid horizontal line indicates the benchmark IGE estimated on longitudinal data (with the dashed lines displaying the 95% confidence interval).

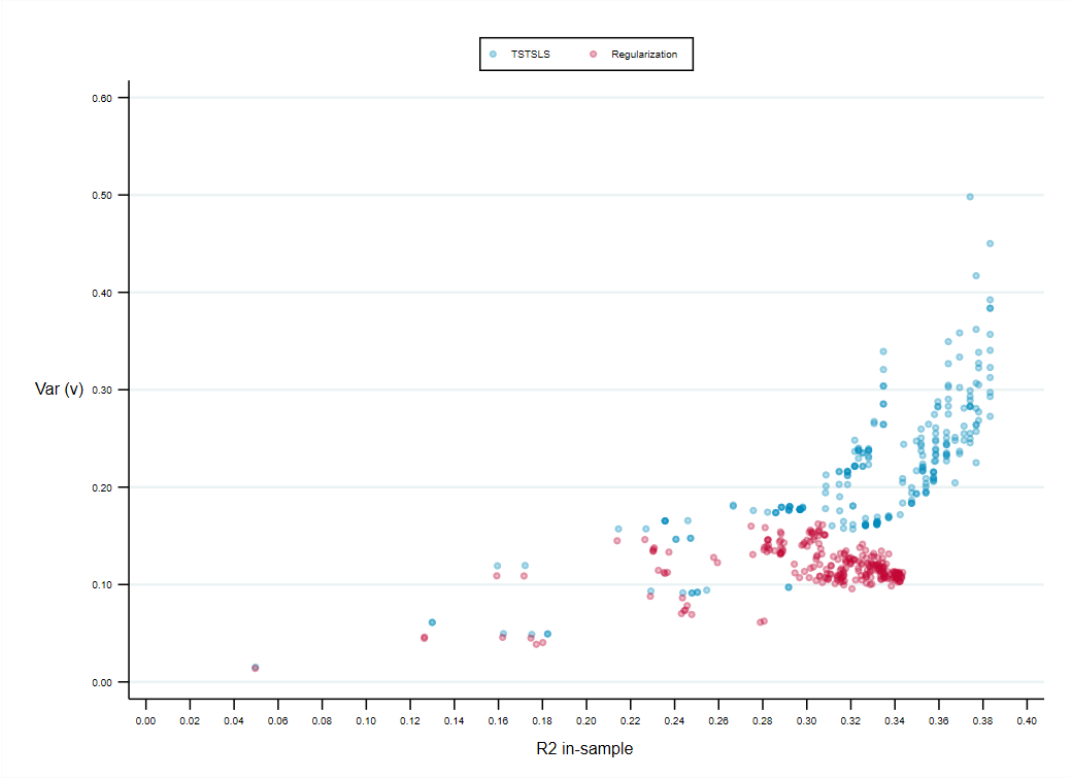
One implication from Figure 1 is that our method improves the prediction of unseen fathers' income for models with a high number of first-stage regressors. While some existing studies in the literature warn against the use of a high number of variables in the prediction equation, this is often motivated by a presumed risk of an increase in the upward bias of the resulting IGE estimate. We show in Figure 2, however, that this presumption may not be correct. The figure plots the relationship between the in-sample R^2 (x-axis) and the IGE (y-axis). It shows that underfitted models (left) tend to produce upwardly biased estimates (even more so when regularized). As the complexity of the model increases, the regularized models tend to converge

to the benchmark longitudinal IGE estimated on real fathers (black solid horizontal line). Instead, the overfitting in the standard TSTSLS (right side of the graph) induces a clear downward bias. The intuition behind this result is that for very imprecise (out-of-sample) models the information embedded in the predicted father's income is so noisy that it attenuates the estimated intergenerational income association. Our second finding is thus that as models become more complex, regularization corrects the downward bias in the IGE.

Figure 3 below provides an explanation for this finding. For more complex models, $var(v_i)$ increases exponentially in the non-regularized models. This leads to a progressively smaller θ in Equation (6), which implies a more severe attenuation bias. In other words, the standard TSTSLS faces a trade-off between the potentially valuable information contained in a large number of regressors with the risk of overfitting the data. Regularization bounds this source of bias, while at the same time trying to extract the useful variation in all possible predictors of parental income.

Figure 3:

$Var(v_i)$ and in-sample R^2



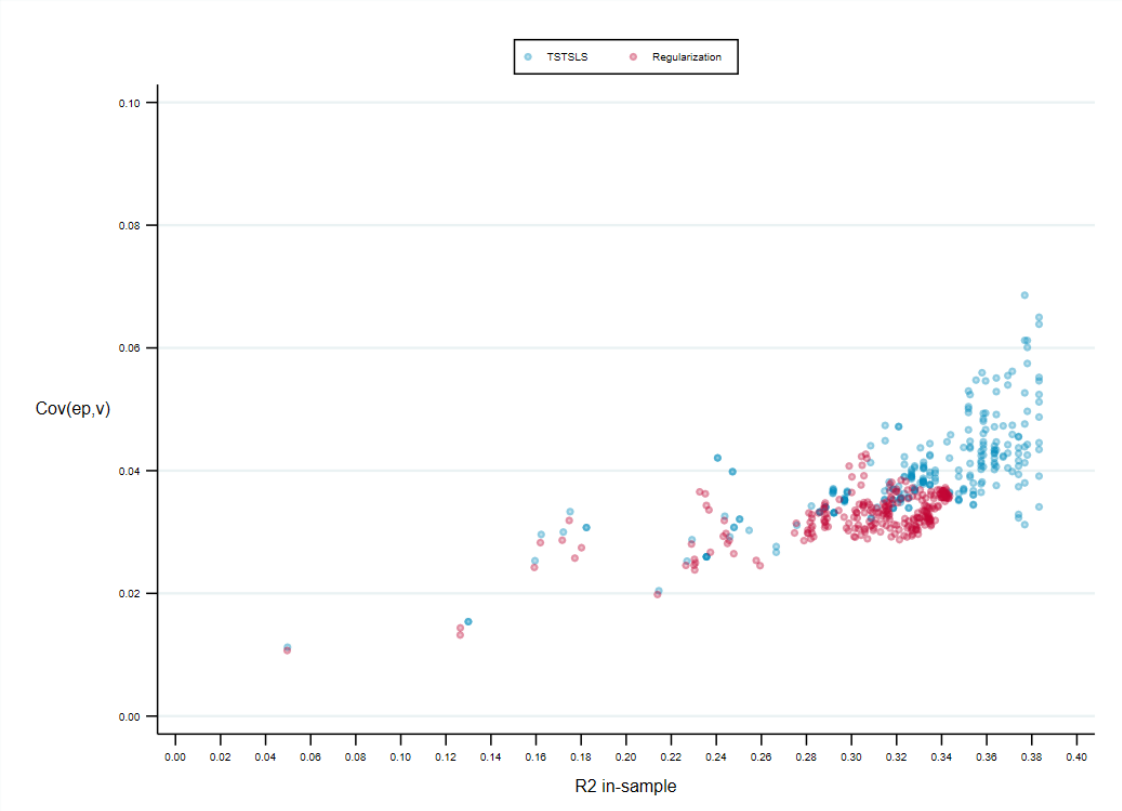
Source: PSID (1982-2011)
Notes: The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the variance component for both standard TSTSLS (blue) and regularized (red) models.

Figure 4 shows that something similar may be happening with respect to the second source of bias in the TSTSLS. As models become more complex, $cov(\epsilon_i, v_i)$ increases. Since

this is one of the drivers of bias *(ii)*, the standard approach once again faces a trade-off between using the potentially valuable information in a larger number of regressors and the risk of a greater bias. Unlike the previous figure, however, here the risk is towards an upward bias from the direct effect of first-stage variables on sons' income. Regularization limits this risk by using a specification of the first-stage model that reduces the residual variation entering directly in the second-stage equation. In other words, by virtue of focusing on the maximum predictive power of the first-stage, the algorithm leaves less room for the included variables to 'bypass' parental income, which bounds the upward bias in the TSTSLS.

Figure 4:

$cov(\epsilon_i, v_i)$ and in-sample R^2



Source: PSID (1982-2011)
Notes: The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the covariance component for both standard TSTSLS (blue) and regularized (red) models.

Table 1 presents IGE estimates for the United States and the corresponding in- and out-of-sample R^2. The first row reports the benchmark IGE estimated on the longitudinal PSID sample linking sons to their real fathers. The estimated value is 0.492, which is consistent with many of the existing estimates of intergenerational income mobility available for the U.S. (Corak, 2013). The second row presents the IGE resulting from the TSTSLS specification that

minimizes the out-of-sample MSE.[7] The IGE estimated by this model is equal to 0.496, which is remarkably close to the one obtained from the longitudinal sample. This result holds when considering the average estimates across the top-5 and top-10 performing models (rows 3 and 4). Overall, Table 1 suggests that by bounding both sources of bias in the TSTSLS, regularization leads to a bias *(i)* and bias *(ii)* of comparable magnitudes. As they operate in different directions, this results in an IGE estimate close to the benchmark.

Table 1:
IGE estimates: Regularization

	IGE	s.e.	First-Stage R^2 (out-of-sample)	First-Stage R^2 (in-sample)	γ	$var(v_i)$	$cov(\epsilon_i, v_i)$	Bias *(i)*	Bias *(ii)*	Final Bias
1. Benchmark (OLS)	0.492	0.062								
2. 'Best' model	0.496	(0.078)	0.261	0.324	0.363	0.124	0.032	-0.200	0.204	0.004
3. Average of top 5 performing models (out-of-sample)	0.487	(0.074)	0.260	0.317	0.373	0.129	0.032	-0.203	0.198	-0.005
4. Average of top 10 performing models (out-of-sample)	0.494	(0.080)	0.260	0.319	0.369	0.127	0.032	-0.200	0.202	0.002
Sample size	1,061	1,061	1,860	1,860	1,061	1,061	1,061	1,061	1,061	1,061

Source: PSID (1982-2011)
Notes: Bootstrapped standard errors (reps 500) in parentheses.

Table 2 shows the estimated levels of intergenerational mobility in the United States using different combinations of first-stage variables for the standard TSTSLS method. The first row reports again the benchmark IGE estimated on the longitudinal PSID sample linking sons to their real fathers. The remaining rows confirm that more complex models tend to underestimate the IGE by increasing the attenuation bias. In particular, the results in the table confirm that it is not advisable to use all the available variables without regularization (row 6). This is because a higher R^2 does not necessarily decrease the bias. In fact, beyond a certain threshold, the attenuation bias becomes substantial. On the other hand, when using only education as predictor of parental income, the IGE suffers from a considerable upward bias. This is due to a combination of low γ and low residual variability in the first-stage model.

It is worth noting that the specification using education and occupation (row 3) delivers an IGE that is fairly close to the longitudinal benchmark. Since this is a common specification

[7] This model includes 164 first-stage regressors and is regularized by an elastic-net with $\lambda = 0.4015$ and $\alpha = 0.0101$. Figure A1 in the Appendix shows the model selection, while the first-stage coefficients are presented in Table A1.

choice in the literature, we may be tempted to interpret this result as a reassuring finding for the reliability of existing estimates. However, it is not possible to know a priori which combination of first-stage predictors delivers the least biased estimate. While this specification appears to be the best in this U.S. sample, it may not be true in other contexts or even in other U.S. samples where this information is reported on a different number of categories or using a different classification. The advantage of using our approach is that it does not require researchers to know *ex ante* the best set of first-stage predictors.

Table 2:
IGE estimates: Standard TSTSLS

	IGE	s.e.	First-Stage R^2 (out-of-sample)	First-Stage R^2 (in-sample)	γ	$var(v_i)$	$cov(\epsilon_i, v_i)$	Bias *(i)*	Bias *(ii)*	Final Bias
1. Benchmark (OLS)	0.492	0.062								
2. Education only	0.929	(0.127)	0.151	0.162	0.226	0.050	0.030	-0.035	0.472	0.437
3. Education + occupation	0.478	(0.073)	0.202	0.222	0.368	0.139	0.037	-0.224	0.210	-0.015
4. Education + occupation + industry	0.379	(0.069)	0.250	0.276	0.412	0.176	0.031	-0.254	0.142	-0.113
5. Education + occupation + industry + race	0.400	(0.068)	0.255	0.282	0.429	0.174	0.034	-0.247	0.154	-0.092
6. Education + occupation + industry + race + interactions	0.219	(0.053)	0.096	0.383	0.455	0.450	0.052	-0.378	0.104	-0.274
Sample size	1,061	1,061	1,860	1,860	1,061	1,061	1,061	1,061	1,061	1,061

Source: PSID (1982-2011)
Notes: Bootstrapped standard errors (reps 500) in parentheses.

Overall, the results in Table 1 and 2 show that the elastic net can limit the risk of bias in the TSTSLS. By bounding the two main sources of bias, which work in opposite ways, the regularization lowers the risk of the estimator moving excessively in either direction. As our approach lets the data find the optimal specification for predicting parental income for any context or data availability, it is no longer necessary to defend arbitrary specifications. This has important consequences for the comparability of IGE estimates across countries and time periods, where the data generating processes are likely to be very different.

4.2 Standard and regularized TSTSLS on sub-optimal data

The previous section highlights the usefulness of our proposed method in a data scenario where we can have a benchmark OLS estimate on longitudinal information. For most countries, however, scholars have access to sub-optimal data sources and cannot estimate the IGE on an intergenerationally-linked sample. These are precisely the situations where our method can be

most valuable, by providing a non-arbitrary criterion to obtain an IGE estimate. We illustrate here an application of our approach on data from an emerging country where long-span income information covering two generations is not available. This represents a common data condition for the developing world, as well as for historical records.

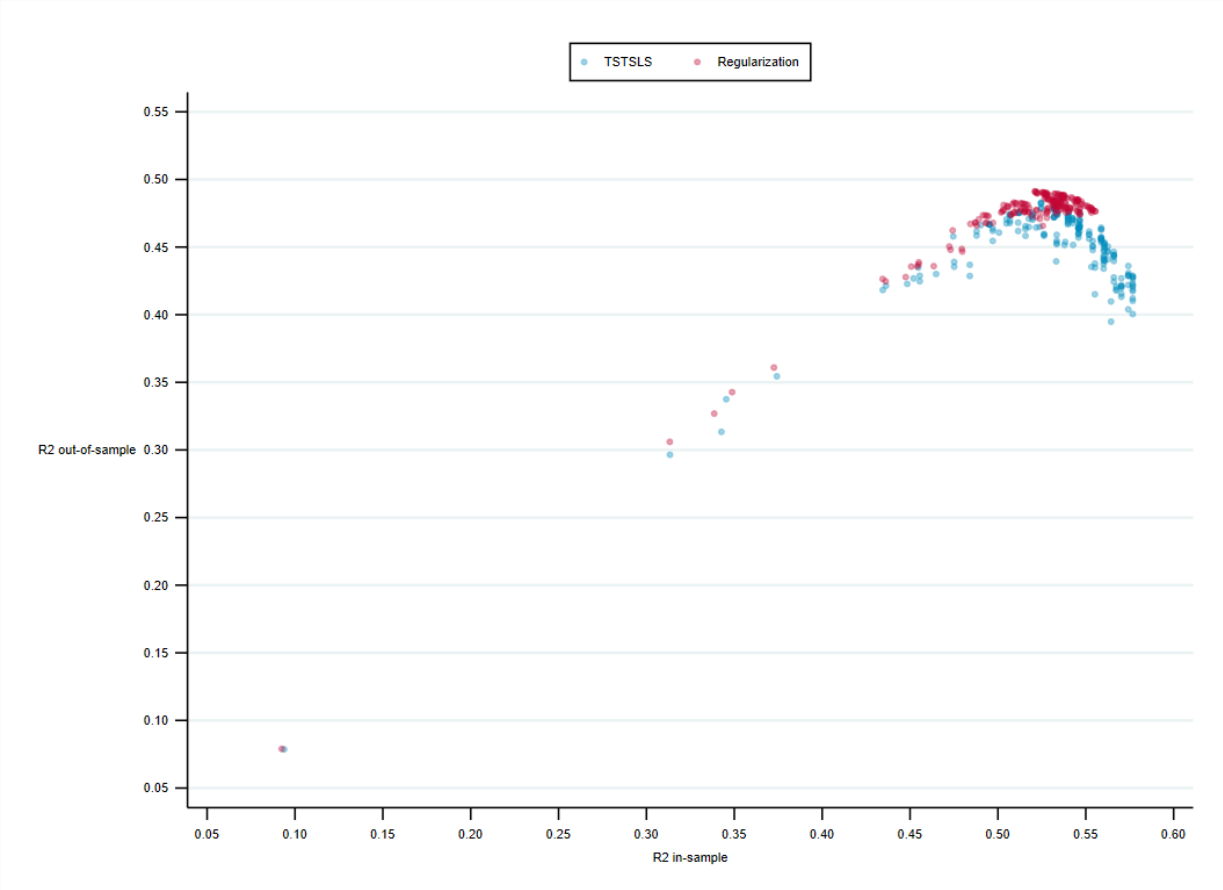
We replicate part of the empirical analysis in the previous section using survey data from South Africa. For simplicity, we use the same data and sample selection rules as in Piraino (2015), who estimates the standard β_{TSTSLS} on the basis of two nationally representative samples.[8] The main sample of 1,241 sons derives from pooling the 2008 to 2012 waves of the National Income Dynamics Study (NIDS), which includes a dedicated section with retrospective information about the parents of respondents. The auxiliary sample of 1,292 pseudo-fathers is based on the Project for Statistics on Living Standards and Development (PSLSD), the first nationally representative survey conducted in South Africa. We use monthly gross employment income, constructed as the sum of wages, salary bonuses, shares of profit, income from agricultural activities, casual and self-employment income. We restrict the analysis to male workers aged 20 to 44 with positive earnings. The first-stage variables used to predict fathers' income are dummies for education (6), occupation (6), province (9), and race (4), plus all pairwise interactions. We thus obtain 1,023 different models and 203 models of varying complexity (i.e. number of regressors).

Figure 5 and 6 use South African data to replicate the analysis in Figure 1 and 2 for the United States. Figure 5 confirms that the non-regularized regression (blue dots) overfits the data for models including a high number of regressors. The pattern is very similar to the one obtained on the U.S. data, showing the decrease in the ability to correctly predict out-of-sample for specifications delivering a very high in-sample R^2. Once again, the elastic-net models (red dots) are able to avoid overfitting, confirming that regularization improves out-of-sample prediction for complex models.

[8] The main difference with respect to the selection rules adopted by Piraino (2015) is that we do not allow the samples to vary across different first-stage specifications according to missing information in the included variables. We use, instead, constant sample sizes of sons and pseudo-fathers across different models.

Figure 5:

Increasing complexity of the first-stage model and predictive performance of standard TSTSLS vs regularized regression: South Africa



Source: PSLSD (1993)
Notes: The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the out-of-sample R^2 estimated by 5-fold cross-validation for both the standard TSTSLS (blue) and regularized (red) models.

Figure 6 shows that the overfitting in the standard TSTSLS results in lower estimated IGEs. Once again, this result is similar to the finding for the United States confirming the intuition that for very imprecise (out-of-sample) models the noisiness in predicted father's income attenuates the estimated intergenerational income association. The regularized regression (red dots) corrects this attenuation bias and stabilizes the IGE as models become more complex. While we cannot estimate $var(v_i)$ and $cov(\epsilon_i, v_i)$ on the South African data, we can assume that the patterns shown in Figure 3 and 4 for the U.S. would extend to this sample. That is, $var(v_i)$ would increase exponentially in the non-regularized models, leading to a progressively more severe attenuation bias (smaller θ in Eq. 6). At the same time, as models become more complex, $cov(\epsilon_i, v_i)$ may increase, leading to an upward bias in the IGE. Regularization can bound both sources of bias. It extracts the variation in all possible first-

stage predictors while limiting the risk of overfitting the data and reducing the extent to which first-stage predictors enter directly in the child income equation.

Figure 6:

In-sample R^2 and estimated IGE for standard TSTLS and regularized models: South Africa

Source: NIDS (2008-2012)
Notes: The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the corresponding IGE estimate for both standard TSTSLS (blue) and regularized (red) models.

Table 3 reports the TSTSLS intergenerational mobility estimates for South Africa along with the corresponding in- and out-of-sample first-stage R^2. Panel A reports the IGE resulting from the TSTSLS specification that minimizes the out-of-sample MSE (row 1) and the average estimates across the top-5 and top-10 performing models (rows 2 and 3). The estimated IGE in these specifications ranges from 0.691 to 0.695.[9] These values are consistent with the evidence from previous studies of South Africa (Piraino, 2015, Finn et al. 2017), which find very low levels of intergenerational mobility.

[9] The best model includes 67 first-stage regressors and is regularized by an elastic-net with $\lambda = 0.123$ and $\alpha = 0.030$.

Table 3:
IGE estimates for South Africa: Regularization vs. Standard TSTSLS

	IGE	s.e.	First-Stage R^2 (out-of-sample)	First-Stage R^2 (in-sample)
A. Regularized TSTSLS				
1. 'Best' model	0.691	(0.089)	0.491	0.521
2. Average of top 5 performing models (out-of-sample)	0.694	(0.088)	0.491	0.522
3. Average of top 10 performing models (out-of-sample)	0.695	(0.087)	0.491	0.523
B. Standard TSTSLS				
4. Education only	0.628	(0.073)	0.337	0.345
5. Education + occupation	0.642	(0.083)	0.421	0.436
6. Education + occupation + province	0.676	(0.071)	0.435	0.455
7. Education + occupation + province + race	0.762	(0.069)	0.466	0.499
8. Education + occupation + province + race + interactions	0.452	(0.072)	0.417	0.577
Sample size	1,241	1,241	1,292	1,292

Source: NIDS (2008-2012) and PSLSD (1993).
Notes: Bootstrapped standard errors (reps 100) in parentheses.

Panel B of Table 3 displays the estimated IGEs using different combinations of first-stage variables for the standard TSTSLS method. Similar to the evidence from the U.S., the most complex model (row 8), which includes all available predictors and their interactions, has the highest in-sample R^2 while delivering a very low IGE as a result of severe attenuation bias. This confirms that a higher R^2 does not necessarily decrease the bias in the TSTSLS estimates. Note also that different combinations of first-stage predictors result in varying IGEs, with the estimates not following the same pattern observed in the United States. This highlights that using similar variables to predict parents' income in different contexts need not have the same effect on the bias of the TSTSLS estimator. Using an objective and data-driven criterion to choose the first-stage specification may thus be preferable to choosing arbitrary combinations and may help increase comparability across countries.

5. Concluding remarks

We suggest a modification to the standard two-sample two-stage approach for estimating the intergenerational income elasticity in sub-optimal data conditions. Our method minimizes the out-of-sample prediction error in the first-stage equation, which provides an objective criterion for choosing across different specifications of the parental income prediction. Using longitudinal data from the United States, we show that our approach decreases the risk of overfitting in the prediction of parental income, while at the same time reducing the potential for an upward bias in the IGE. Importantly, our two-sample estimates converge to the benchmark IGE estimate from longitudinal data. We replicate part of the analysis on South African data and find consistent results. Overall, the empirical evidence in the paper suggests that a simple machine learning method may improve the reliability and comparability of intergenerational mobility estimates for a large section of the world's population.

References

Aaronson, D. and B. Mazumder. (2008). "Intergenerational economic mobility in the United States, 1940 to 2000". *Journal of Human Resources*, 43(1), 139–172.

Arlot, S. and A. Celisse. (2010). "A survey of cross-validation procedures for model selection." *Statistics surveys* 4 (2010): 40–79.

Björklund, A. and M. Jäntti. (1997). "Intergenerational income mobility in Sweden compared to the United States." *American Economic Review,* 87(4): 1009–1018.

Björklund, A. and M. Jäntti. (2009). 'Intergenerational Income Mobility and the Role of Family Background'. In W. Salverda, B. Nolan, and T. Smeeding, Tim (eds), *Handbook of Economic Inequality*. Oxford: Oxford University Press.

Blanden, J. (2013). "Cross- country rankings in intergenerational mobility: a comparison of approaches from economics and sociology". *Journal of Economic Surveys*, 27(1), 38-73.

Chetty, R., N. Hendren, P. Kline, and E. Saez. (2014). "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *Quarterly Journal of Economics*, 129 (4), 1553–1623.

Brunori, P., F.H.G. Ferreira, V. Peragine, P. Piraino, R. Van der Weide, F. Bloise, R. Gupta, L. Gasparini, C. Lakner, F. Luppi, D. Mahler, A. Narayan, G. Neidhöfer, F. Palmisano, T. Randazzo, T. Rampino, L. Serlenga, J. Serrano, M. Triventi. (2020) "Equal chances: equality of opportunity and intergenerational mobility around the world", University of Bari, *mimeo*.

Clark, G. (2014). *The Son Also Rises: Surnames and the History of Social Mobility*. Princeton University Press.

Corak, M. (2006). "Do poor children become poor adults? Lessons from a cross-country comparison of generational earnings mobility." *Research on economic inequality*, 13(1), 143-188.

Corak, M. (2013). "Income inequality, equality of opportunity, and intergenerational mobility", *Journal of Economic Perspectives*, 27(3): 79–102.

Emran, M.S. and F.J. Shilpi. (2019). "Economic approach to intergenerational mobility: Measures, methods, and challenges in developing countries". (No. 2019/98). UNU-WIDER Working Paper.

Finn, A., Leibbrandt, M., and Ranchhod, V. (2017). Patterns of persistence: Intergenerational mobility and education in South Africa. Cape Town: SALDRU, UCT. (SALDRU Working Paper Number 175/ NIDS Discussion Paper 2016/2).

Gong, H., A. Leigh, and X. Meng. (2012). "Intergenerational income mobility in urban China." *The Review of Income and Wealth*, 58(3), 481–503.

Haider, S. and G. Solon. (2006). "Life-cycle variation in the association between current and lifetime earnings". *American Economic Review*, 96(4), 1308-1320.

James, G., Witten D., Hastie T. and Tibshirani R. (2013). "An introduction to statistical learning". Springer, New York.

Jerrim, J., A. Choi, and R. Simancas. (2016). "Two-Sample Two-Stage Least Squares (TSTSLS) estimates of earnings mobility: how consistent are they?" *Survey Research Methods*. Vol. 10, No. 2, pp. 85–102.

Narayan, A., R. Van der Weide, A. Cojocaru, C. Lakner, S. Redaelli, D.G. Mahler, R.G. Ramasubbaiah, and S. Thewissen. (2018). Fair Progress?: *Economic Mobility Across Generations Around the World*. World Bank Group.

Nybom, M. and J. Stuhler. (2016). "Heterogeneous income profiles and lifecycle bias in intergenerational mobility estimation." *Journal of Human Resources*, 51(1), 239-268.

Olivetti, C. and D. Paserman. (2015). "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940." *American Economic Review*, 105 (8): 2695-2724.

Piraino, P. (2015). "Intergenerational Earnings Mobility and Equality of Opportunity in South Africa." *World Development,* 67: 396–405.

Santavirta, T., and J. Stuhler. (2019). "Name-Based Estimators of Intergenerational Mobility: Evidence from Finnish Veterans." *mimeo.*

Solon, G. (1992). "Intergenerational income mobility in the United States." *American Economic Review*, 82(3), 393–408.

Solon, G. (2002). "Cross-country differences in intergenerational earnings mobility". *Journal of Economic Perspectives*, 16(3), 59-66.

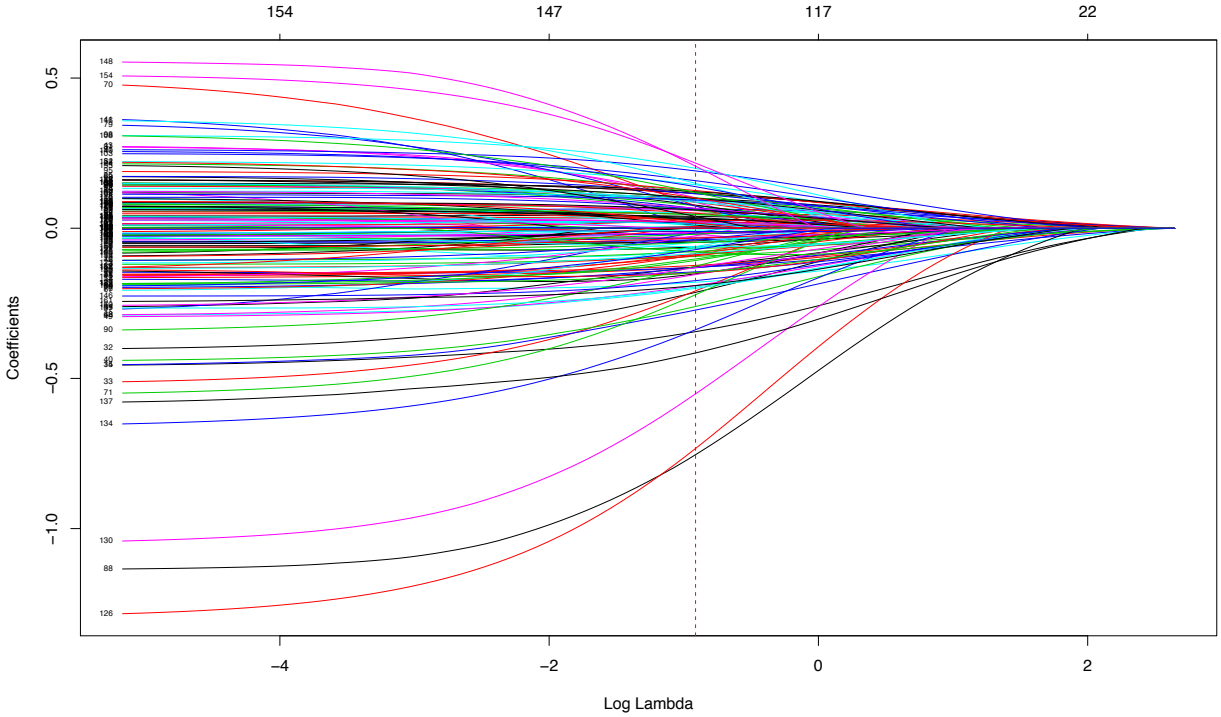
Zou, H. and T. Hastie. (2015). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2: 301–320."

Supplemental Appendix

The model selection by the elastic-net on the U.S sample is represented below. Tuning the model by 5-fold cross validation, we test 100 λ_s and 100 α_s. Each line represents the value of a coefficient for different values of λ (in logs) for the selected α. Larger λ shrinks the coefficients toward zero. The values that minimize the MSE out-of-sample are $\lambda = 0.4015$ and $\alpha = 0.0101$.

Figure A1:

Elastic net coefficient selection



Source: PSID (1982).
The red vertical line represents the value of λ that minimize the out-of-sample MSE.

Table A1:

First-stage elastic-net regression (best model)

First-stage predictors	Coefficients
Education:	
0-5 grades (ef1)	-0.067
Grade school (ef2)	-0.160
Some high school (ef3)	-0.128
High school (ef4)	-0.036
12 grades plus non-academic training (ef5)	0.000
Some college, no degree; associate's degree (ef6)	0.037
College BA and no advanced degree mentioned (ef7)	0.113
College, advanced or professional degree (ef8)	0.148
Occupation:	
Legislators, senior officials, and managers (of1)	0.074
Professionals (of2)	0.083
Technicians and associate professionals (of3)	0.039
Clerks (of4)	0.000
Service and sales workers (of5)	-0.009
Skilled agricultural and fishery workers (of6)	-0.424
Craft and trades workers (of7)	0.000
Operators and assemblers (of8)	0.000
Elementary occupations (of9)	-0.166
Race:	
White (rf1)	0.051
African American (rf2)	-0.073
Hispanic (rf3)	0.034
Industry Sector:	
Agriculture, forestry, and fishing (sf1)	-0.225
Mining (sf2)	0.088
Construction (sf3)	-0.059
Manufacturing (sf4)	0.030
Transportation, communications, and public utilities (sf5)	0.060
Wholesale and retail trade (sf6)	-0.039
Finance, insurance, and real estate (sf7)	0.000
Private services (sf8)	-0.021
Public administration (sf9)	0.056
Pairwise interactions (excluding coefficients set to zero by the algorithm):	
ef1*sf3	0.116
ef1*sf4	-0.256
ef1*sf5	-0.261
ef2*sf1	-0.278
ef2*sf5	-0.101
ef2*sf6	-0.281
ef2*sf7	0.065
ef2*sf8	-0.212
ef2*sf9	0.119
ef3*sf2	-0.097
ef3*sf5	0.021
ef3*sf6	-0.196
ef3*sf7	-0.129
ef3*sf8	-0.119
ef3*sf9	0.088
ef4*sf1	-0.125
ef4*sf2	0.140
ef4*sf3	-0.015
ef4*sf5	0.029
ef4*sf6	-0.098
ef4*sf7	0.038
ef4*sf8	-0.149
ef5*sf5	0.076
ef5*sf6	0.077
ef5*sf7	-0.066
ef5*sf8	-0.177
ef5*sf9	0.076
ef6*sf1	0.097
ef6*sf2	-0.299
ef6*sf3	-0.042
ef6*sf5	0.087
ef6*sf6	0.021
ef6*sf7	-0.108
ef7*sf1	0.110
ef7*sf4	0.117
ef7*sf6	0.129

ef7*sf7	0.126
ef8*sf1	-0.839
ef8*sf3	-0.162
ef8*sf4	0.189
ef8*sf8	0.132
of1*sf2	0.228
of1*sf4	0.137
of1*sf7	0.174
of1*sf8	0.099
of2*sf3	0.142
of2*sf4	0.094
of2*sf6	-0.056
of3*sf3	0.168
of3*sf5	0.112
of4*sf4	-0.045
of4*sf7	-0.847
of4*sf8	-0.104
of4*sf9	0.079
of5*sf3	-0.640
of5*sf4	0.041
of5*sf5	0.111
of5*sf7	-0.389
of5*sf8	-0.208
of6*sf1	-0.392
of7*sf1	-0.124
of7*sf2	0.165
of7*sf3	-0.092
of7*sf4	0.087
of7*sf5	0.073
of7*sf8	-0.177
of8*sf1	0.244
of8*sf2	-0.016
of8*sf4	-0.066
of8*sf6	-0.069
of8*sf7	0.241
of8*sf8	0.051
of9*sf5	-0.115
of9*sf6	-0.169
Constant	-0.030
R-squared in-sample	0.324
Cross-validation R-squared	0.261
Alpha	0.010
Lambda	0.402
Nr. of folds	5
Nr. of alpha tested	100
Nr. of lambda tested	100
Sample size	1,860

Source: PSID (1982)

Table A1 reports the first-stage coefficients of the estimated elastic net in the U.S. sample. The shrunken coefficients are consistent with what one would expect for a large majority of the controls. Education up to high school (ef3) has a negative sign and is positive and increasing for higher levels of education. Occupation variables also have the expected sign: service and sales workers, skilled agricultural and fishery workers, and elementary occupations have a negative sign, all other occupations have a positive sign (professional has the largest coefficient), but clerks, craft and trades workers, and operators and assemblers have a coefficient shrunken very close to zero. As far as race is concerned, the African American dummy has a negative coefficient. Coefficients estimated for industry sector and for the interacted variables are less straightforward to interpret but some have a non-negligible magnitude.