The Journal of Economic Inequality (2021) 19:643–665
https://doi.org/10.1007/s10888-021-09495-6

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

Francesco Bloise[1] · Paolo Brunori[2] · Patrizio Piraino[3]

Received: 26 March 2020 / Accepted: 17 May 2021/ Published online: 13 August 2021

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 lower-income countries and for estimating long-run trends across multiple generations and historical periods. We propose applying machine learning methods to improve the reliability and comparability of such estimates. Supervised learning algorithms minimize the out-of-sample prediction error in the parental income imputation and provide an objective criterion for choosing across different specifications of the first-stage equation. We use our approach on 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 income mobility · Machine learning · Two-sample two-stage least squares

JEL classification J62 · D31 · D63

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

✉ Francesco Bloise
francesco.bloise@uniroma3.it

Paolo Brunori
paolo.brunori@unifi.it

Patrizio Piraino
ppiraino@nd.edu

[1] University of Roma Tre, Rome, Italy

[2] London School of Economics, London, UK

[3] University of Notre Dame, Notre Dame, IN, USA

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 applying a machine learning approach to 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 low and middle-income nations (Narayan et al. 2018; Brunori et al. 2020).

Machine learning (ML) methods are increasingly integrated into the statistical toolkit of economists (Athey and Imbens 2017; Belloni et al. 2014; McKenzie and Sansone 2019; Mullainathan and Spiess 2017; Varian 2014; Blundell and Risa 2019) have explored the possibility to use ML algorithms to improve our ability to understand the intergenerational transmission of income. They show how a data-driven approach can shed light on generally ignored channels of transmission of income and wealth. We contribute to this debate by showing how ML improves the imputation of parental income in the TSTSLS and provides an objective criterion for choosing across different specifications of the prediction equation. Using supervised learning algorithms, we are able to identify the model specification that

minimizes the out-of-sample prediction error of parental income. Such a criterion is applicable to different data conditions and 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 approach, searching for the specific algorithms that best exploit the information embedded in all available predictors of parental income. We consider a number of algorithms to minimize the out-of-sample prediction error and compare their relative predictive ability. Based on this exercise, we opt for a shrinkage method (Zou and Hastie 2005; Meinshausen 2007), which avoids overfitting by shrinking the standard linear regression coefficients. While the choice of the algorithm is based on its predictive performance, an attractive aspect of regularized regression is that it improves the accuracy 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 the Panel Study of Income Dynamics (PSID), a longitudinal income survey data from the United States. 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 Stuhler 2020) 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 variability of 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 ML 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:

$$y_i^c = \alpha + \beta y_i^p + \epsilon_i \tag{1}$$

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 lower-income 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:

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

where y_{it}^{ps} is the log 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:

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

where y_i^c is the log income of children. $\widehat{y}_i^p = \widehat{\phi} z_i^p$ is the imputed log income of *unseen* parents,[2] and z_i^p are recall variables analogous to z_i^{ps}. Note that Eq. (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.[3]

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 Eq. (3) and the elasticity estimated on Eq. (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] To calculate standard errors when using predicted income, we use the bootstrap procedure (see also Björklund and Jäntti, 1997).

[3] In fact, Eq. (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 $\widehat{y}_i^p$ on y_i^p:

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

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 Eq. (4) to express the probability limit of the TSTSLS estimator as follows:

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

Which, using Eq. (1), can be rewritten as

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

where $\theta = \frac{\gamma var\left(y_i^p\right)}{\gamma^2 var\left(y_i^p\right) + var\left(v_i\right)}$ represents bias (*i*), and the ratio $\frac{cov(\epsilon_i, v_i)}{\gamma^2 var\left(y_i^p\right) + var\left(v_i\right)}$ 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.[4] 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

[4] The term describing bias *(i)* above is similar to the first term in Eq. (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.

describe the machine learning approach used to minimize the out-of-sample prediction error in the parental income imputation (Eq. 2).

3 Method

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

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

where y_0^p is the income of the real parent of individual *0* (a person we do not observe) and $\widehat{f}\left(z_0^{ps}\right)$ 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:

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

where $var\left(\widehat{f}\left(z_0^{ps}\right)\right)=\mathbb{E}\left\{\left[\widehat{f}\left(z_0^{ps}\right)-\mathbb{E}\left(\widehat{f}\left(z_0^{ps}\right)\right)\right]^2\right\}$ 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: $\left[f\left(z_0^{ps}\right)-\mathbb{E}\left(\widehat{f}\left(z_0^{ps}\right)\right)\right]^2$ is the squared *bias* of the model, which quantifies the error that is introduced by approximating an unknown data generating process by a simpler model (for 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 Eq. (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. Our prediction problem has both a relatively low number of observations and a relatively low number of predictors. While we are not dealing with 'big data'—the typical environment where ML algorithms outperform standard econometric models— there is a rich range of ML algorithms that can perform well in such contexts (Varian 2014; Hastie et al. 2009). The first step in our empirical analysis it to use the approach proposed by Mullainathan and Spiess (2017) to compare the predictive accuracy of the following models: OLS, Ridge regression, LASSO, relaxed LASSO, Elastic net, Boosted regression, and Random forests. The results of this exercise are reported in Appendix A, where we show that relaxed least absolute shrinkage and selection operator (relaxed LASSO) outperforms the other methods.[5] We thus estimate the first-stage regression in the TSTSLS by implementing a version of the relaxed LASSO operator introduced by Meinshausen (2007).

Let us first consider the elastic-net shrinkage operator introduced by Zou and Hastie (2005). An elastic-net obtains the regression coefficients by minimizing:

[5] We also tested neural network, a popular algorithm when dealing with big data. For the sake of brevity, we do not report the results in Appendix A because our preliminary analysis using this algorithm led to very poor predictive performance in our context.

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

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 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.

The elastic net performs both variable selection and coefficient shrinkage. When $\lambda > 0$ coefficients are shrunk toward zero, and when λ and α are sufficiently large all components of the prediction function are set to zero.[6] However, it has been shown that the use of the same parameter (λ) to perform both variable selection and coefficients' shrinkage can be less effective than using two separate parameters (Efron et al. 2004). As a way to address this issue, Meinshausen (2007) introduces a modification of the standard LASSO called 'relaxed LASSO' that uses two parameters (λ and ϕ). The algorithm can be understood as proceeding in two steps: first a subsample of regressors is selected by estimating a LASSO, then a regularization is performed on the coefficients of a regression that includes only the variables selected in the first step.

The relaxed LASSO approach proposed by Meinshausen (2007) can be estimated by using an additional LASSO or a simple OLS in the second step. Other combinations are also possible, which include the elastic-net shrinkage operator (Hastie et al. 2017). In our case, we estimate a relaxed fit by tuning LASSO in the first step and elastic-net in the second.[7] Therefore, the relaxed LASSO minimizes the following penalty function:

$$\sum_{i=1}^{n}\left(y_i - b_o - b_1\delta_1 X_{1,i} - b_2\delta_2 X_{2,i} - b_k\delta_k X_{k,i}\right)^2 + \lambda\phi\left(\alpha\sum_{j=1}^{k}\delta_j|b_j| + (1-\alpha)\sum_{j=1}^{k}\delta_j b_j^2\right) \quad (10)$$

[6] To have an intuition of why shrinkage improves prediction relative to OLS consider the case of an overfitted model that contains too many regressors. Generally, in an overfitted model all coefficients differ from zero. Coefficients that have no predictive power will have a value around zero but their exact value will depend on the random sample drawn from the population. The fact that coefficients vary with the particular sample observed implies a high variance of the model (first component of Eq. 8) and a low predictive ability out-of-sample. In such situation, the obvious solution consists in selecting a subset of regressors. While this introduces the risk of a small bias, it will result in a substantial drop in the model variance. However, setting all regression coefficients exactly to zero may not be necessary. Some regressors may contain useful information to predict the dependent variable, but such information is not enough to justify the inclusion of the entire coefficient obtained from OLS. Rather than choosing between excluding those regressors altogether or using their OLS coefficients, shrinkage methods decrease the model variance by reducing the coefficients' absolute value.

[7] Since elastic net is a generalization of the other methods, in the event that LASSO or OLS have higher predictive ability in the second step, the elastic net will be equal to a LASSO or an OLS.

where, for every λ, $\delta_j = 1$ if in the LASSO $b_j \neq 0$, and $\delta_j = 0$ otherwise. In practice, after the set of "active" regressors has been determined by estimating a LASSO, their coefficients are shrunk by a different regularization parameter $\lambda\phi$, where$\phi \leq 1$.

Using relaxed LASSO, we obtain different sets of b_s depending on the value of λ, α and ϕ. In statistical learning terminology, this implies that the algorithm needs to be tuned so as to obtain a more accurate model specification. Among all possible specifications, we aim at finding the values λ, α and ϕ so that Eq. (7) is minimized. A standard method for tuning is *k*-folds cross-validation. In this case, at reasonable computational cost, cross-validation provides a direct estimate of the out-of-sample prediction error under very weak assumptions (Arlot and Celisse 2010). A standard strategy is to consider a large number of meaningful values for the three parameters (λ, α and ϕ) and to estimate (7) for all their possible combinations. Appendix A reports how we tune and assess the predictive performance of relaxed LASSO as well as of the other algorithms we considered.

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 literature on lower and middle-income countries.

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.[8] 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.[9]

We then obtain an *auxiliary* sample of 1,860 pseudo-fathers aged 30–60 using the 1982 wave of the PSID.[10] In both the *main* and *auxiliary* samples, we use yearly gross employment

[8] 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 (Haider and Solon 2006; Nybom and Stuhler 2016; Chen et al. 2017).

[9] The international literature on intergenerational mobility shows that there are often differences in IGE estimates by gender and across cohorts. While our method could be extended to mothers and daughters or to a different time period, we see this as a separate contribution which would also require analyzing gender and cohort differentials. This analytical simplification does not impede us to advance the main argument of the paper, which is to show that the regularized TSTSLS performs better than the standard TSTSLS.

[10] We choose the year 1982 to obtain a sample of pseudo-fathers that is more likely to be representative of the sample of real fathers, given that the average year of observation of actual fathers' gross labour income is 1981.5. Below, we check the robustness of our results to using a larger sample of pseudo-fathers.

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 first selects a subset of regressors by estimating a LASSO and then lets the elastic-net find the degree of regularization that minimizes the out-of-sample prediction error based on the selected 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 1,023 different models.[11] Amongst models with an equal number of regressors, we select the one with the highest R^2 (in-sample).[12] 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.

The first noticeable result from Fig. 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 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.

One implication from Fig. 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 Fig. 2, however, that this presumption may not be correct. The figure plots the relationship between model complexity, the in-sample R^2 reported in the 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.

[11] 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).

[12] 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 approach is presented in detail in James et al. (2013), Chap. 6.

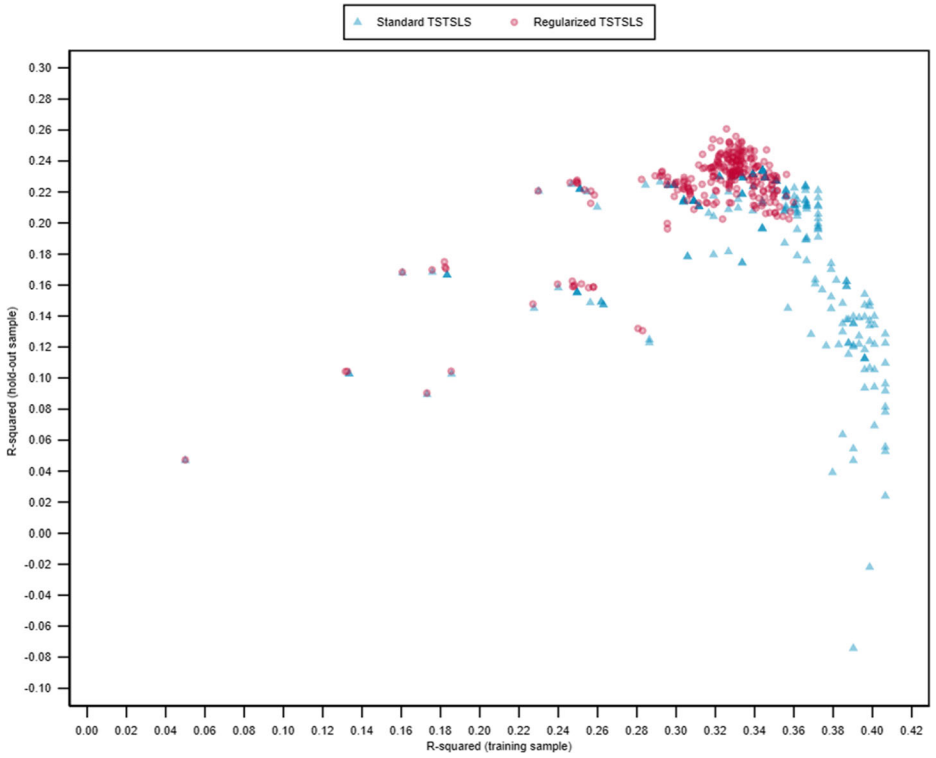

Fig. 1 Increasing complexity of the first-stage model and predictive performance of standard TSTSLS vs. regularized regression. *Source:* PSID (1982). *Notes*: The horizontal axis reports the highest in-sample R^2 for each possible number of regressors (complexity of the model specified). The vertical axis reports the out-of-sample R^2 for both the standard TSTSLS (blue) and regularized (red) models

Figure 3 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 Eq. (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 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.

Table 1 presents the 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

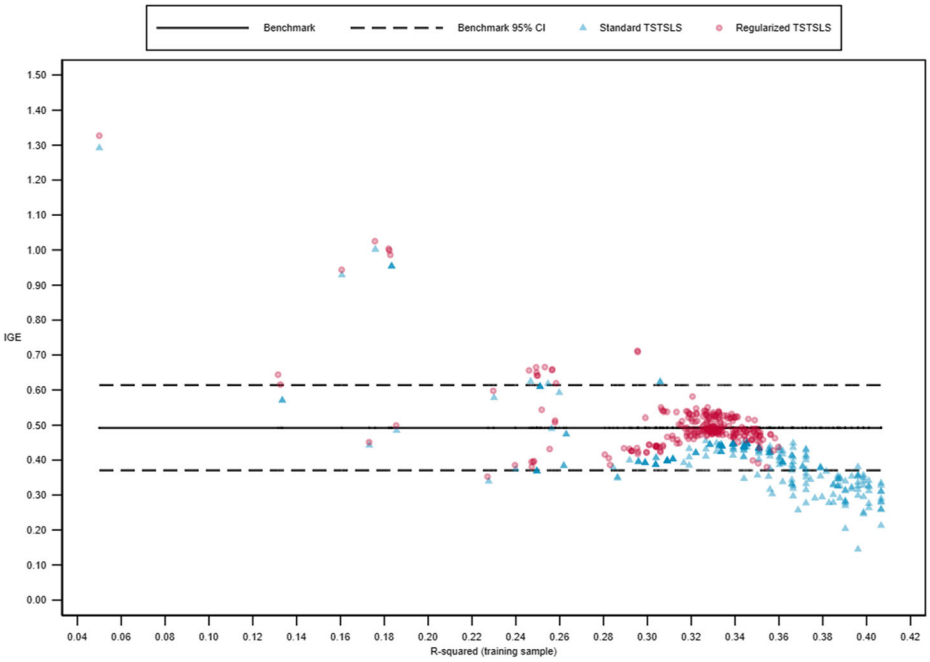

Fig. 2 In-sample R^2 and estimated IGE for standard TSTSLS and regularized model (red). *Source*: OLS benchmark: PSID longitudinal sample (1968–1992) and PSID (2011); standard and regularized TSTSLS: PSID (1982) and PSID (2011). *Notes:* The horizontal axis reports the highest in-sample R^2 for each 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)

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). Rows 2 to 4 display the IGEs resulting from the regularized TSTSLS specifications that minimize the out-of-sample MSE.[13] The IGE estimated by these models range between 0.483 and 0.494, which are remarkably close to the one obtained from the longitudinal sample. This 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.[14]

The results in Panel A of Table 1 also show a substantial stability of the parameters across the top 1, 5 or 10 performing models from the regularized TSTSLS. Averaging across top performing models reduces the sensitivity to random noise in the observed sample, as a single estimate may be more sensitive to the variance in the expected squared error (as shown in Eq. 8). This intuition is

[13] The 'best' model includes 134 first-stage regressors and is regularized in the second step of the relaxed LASSO by an elastic-net with $\lambda = 0.051$ and $\alpha = 0.071$.

[14] For robustness, we re-estimate the regularized IGEs using a larger auxiliary sample. Table 5 in Appendix B presents the estimates obtained using waves 1981, 1982, and 1983 of the PSID. The out-of-sample R^2 increases when we run relaxed LASSO on a larger sample of pseudo-fathers but the estimated IGE remains very close to the benchmark OLS estimate. Regularization thus appears to perform well on small samples, which are the ones usually at disposal of scholars that exploit TSTSLS. We also ran a robustness check where we imputed missing values for the predictors and created a missing indicator following Mullainathan and Spiess (2017). Table 6 in Appendix B shows that the main results do not vary.

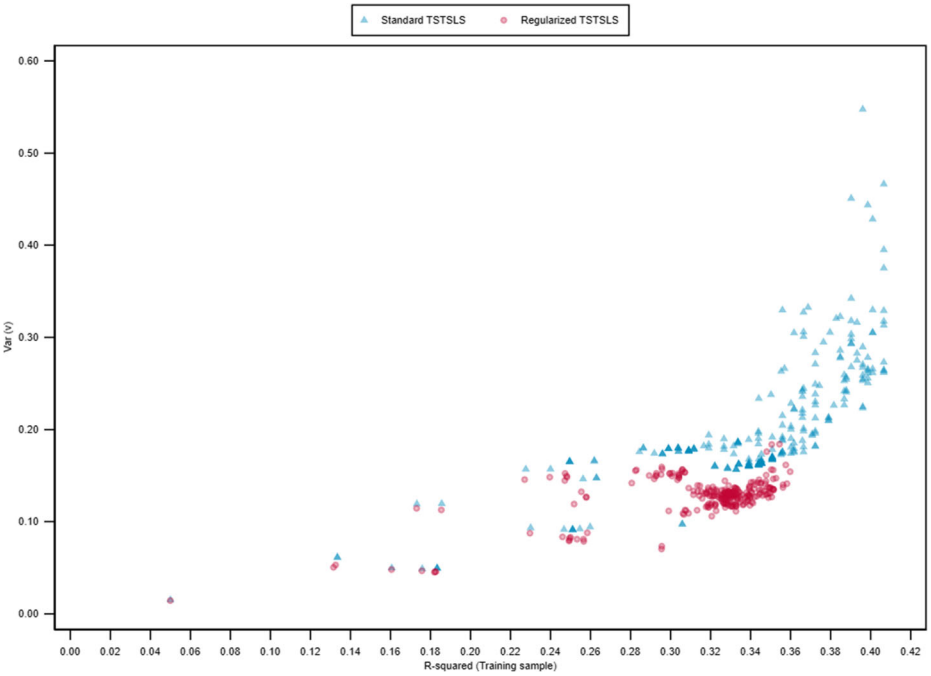

Fig. 3 $\boldsymbol{Var}(\boldsymbol{v_i})$ and in-sample R^2. *Source:* PSID longitudinal sample (1968–1992), PSID (1982) and PSID (2011). *Notes:* The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the variance component for both standard TSTSLS (blue) and regularized (red) models

also confirmed in Figure A1 in Appendix A, where we show the relationship between the out-of-sample R^2 and the IGE estimates for all the ML algorithms we evaluated. The models with the highest out-of-sample prediction accuracy deliver estimated IGEs that are closely clustered around the benchmark value obtained from the longitudinal data.

Panel B of Table 1 shows the estimated levels of intergenerational mobility in the United States using different combinations of first-stage variables for the standard TSTSLS method. Estimates 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 #9). 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.[15]

It is worth noting that the specification using education and occupation (row #6) delivers an IGE that is fairly close to the longitudinal benchmark. Since this is a common specification 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

[15] Note that confidence bounds overlap across some specifications and in certain cases include the benchmark IGE even in if the out-of-sample R^2 is low. This is partly due to the use of conservative standard errors obtained via the bootstrap procedure.

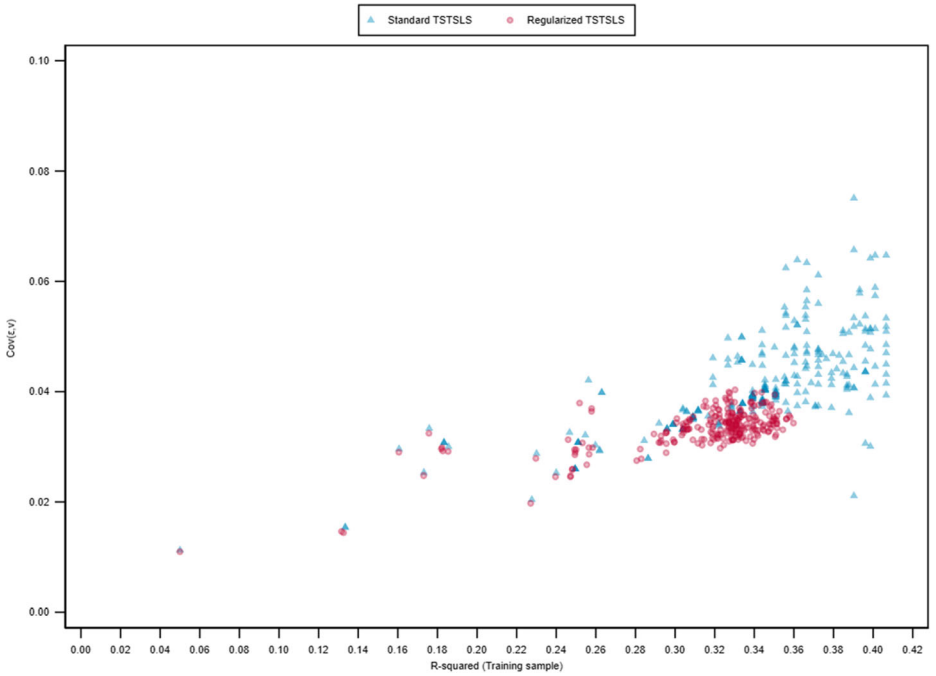

Fig. 4 $cov(\epsilon_i, v_i)$ and in-sample R^2. *Source*: PSID longitudinal sample (1968–1992), PSID (1982) and PSID (2011). *Notes*: The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the covariance component for both standard TSTSLS (blue) and regularized (red) models

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.

Overall, the results in Table 1 show that regularization can limit the risk of bias in the TSTSLS. By bounding the two main sources of bias, which work in opposite ways, 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 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.

Table 1 IGE estimates: Regularization vs. standard TSTSLS

	IGE	First stage R^2 (hold-out sample)	First stage R^2 (training sample)	γ	$var(v_i)$	$cov(\epsilon_i, v_i)$	Bias (i)	Bias (ii)	Final Bias
1. Benchmark (OLS)	0.492								
	(0.062)								
A. Regularized TSTSLS									
2. 'Best' model (hold-out sample)	0.483	0.261	0.326	0.390	0.134	0.034	-0.207	0.198	-0.009
	(0.070)	(0.060)							
3. Average of top 5 performing models (hold-out sample)	0.494	0.255	0.326	0.373	0.129	0.034	-0.206	0.208	0.002
	(0.073)	(0.059)							
4. Average of top 10 performing models (hold-out sample)	0.491	0.253	0.328	0.378	0.131	0.034	-0.206	0.205	-0.002
	(0.072)	(0.060)							
B. Standard TSTSLS									
5. Education only	0.929	0.168	0.161	0.226	0.050	0.030	-0.035	0.472	0.437
	(0.120)	(0.039)							
6. Education + occupation	0.478	0.166	0.232	0.368	0.139	0.037	-0.224	0.210	-0.015
	(0.070)	(0.060)							
7. Education + occupation + industry	0.379	0.225	0.284	0.412	0.176	0.031	-0.254	0.142	-0.113
	(0.064)	(0.066)							
8. Education + occupation + industry + race	0.400	0.227	0.292	0.429	0.174	0.034	-0.247	0.154	-0.092
	(0.064)	(0.067)							
9. Education + occupation + industry + race + interactions	0.213	0.082	0.407	0.450	0.466	0.053	-0.382	0.103	-0.279
	(0.050)	(0.083)							
Sample size	1,061	372	1,488	1,061	1,061	1,061	1,061	1,061	1,061

Source: PSID longitudinal sample (1969–1992) and PSID (2011) for OLS benchmark; PSID (1982) and PSID (2011) for standard and regularized TSTSLS

Notes: Specifications 5–8 include only main effects, specification 9 includes all main effects and all pairwise interactions. Bootstrapped standard errors (reps 200) in parentheses

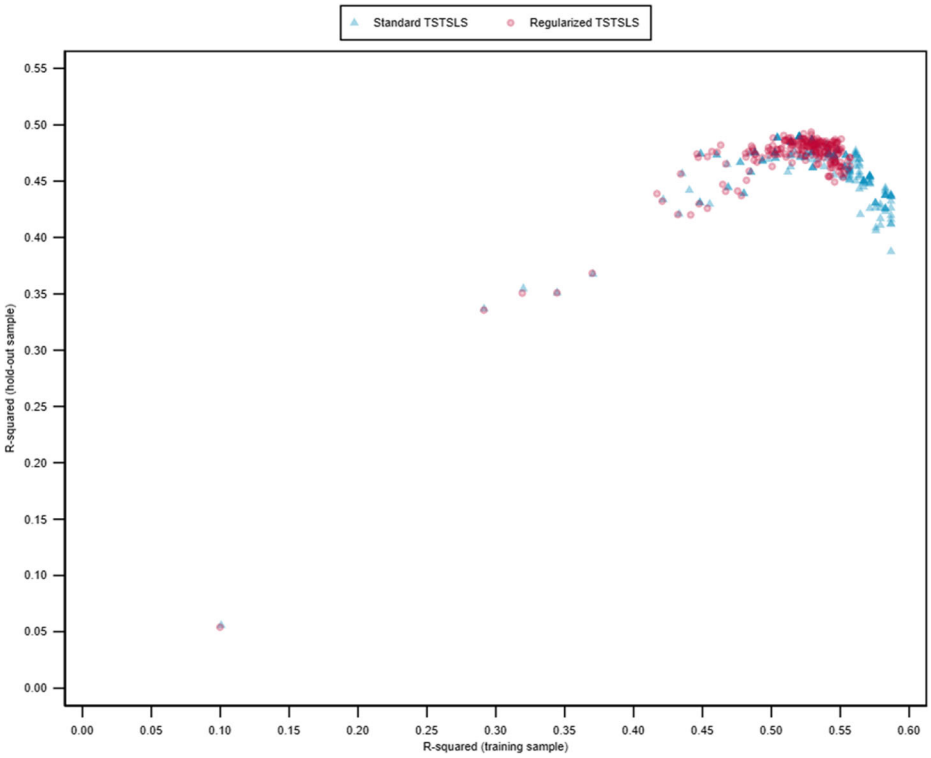

Fig. 5 Increasing complexity of the first-stage model and predictive performance of standard TSTSLS vs. regularized regression: South Africa. *Source:* PSLSD (1993). *Notes:* The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the out-of-sample R^2 for both the standard TSTSLS (blue) and regularized (red) models

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.[16] 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 1993), 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).

Figures 5 and 6 use South African data to replicate the analysis in Figs. 1 and 2 for the United States. Figure 5 confirms that the non-regularized regression (blue dots) overfits the

[16] 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.

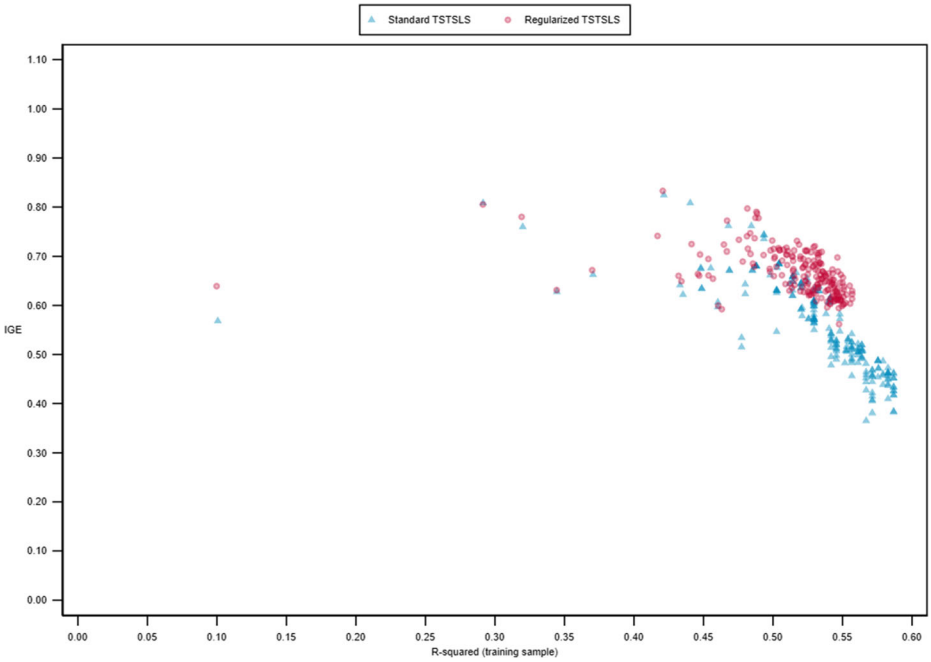

Fig. 6 In-sample R^2 and estimated IGE for standard TSTLS and regularized models: South Africa. *Source:* PSLSD (1993) and NIDS (2008–2012). *Notes:* The horizontal axis reports the highest in-sample R^2 for each possible number of regressors. The vertical axis reports the corresponding IGE estimate for both standard TSTSLS (blue) and regularized (red) models

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 relaxed LASSO (red dots) is able to avoid overfitting, confirming that regularization improves out-of-sample prediction for complex 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 be certain that $var(v_i)$ would increase with complexity in the non-regularized models, leading to a progressively more severe attenuation bias (smaller θ in Eq. 6). Regularization bounds such source of bias.

Table 2 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 regularized TSTSLS specification that minimizes the out-of-sample MSE (row 1) and the average estimates across the top-5 and top-10 performing models in terms of out-of-sample R^2 (rows 2 and 3). The estimated IGE in these specifications range from 0.632 to 0.670.[17] 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.

Panel B of Table 2 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

Table 2 IGE estimates for South Africa: Regularization vs. Standard TSTSLS

	IGE	First-Stage R^2 (hold-out sample)	First-Stage R^2 (training sample)
A. Regularized TSTSLS			
1. 'Best' model	0.632	0.494	0.529
	(0.091)	(0.048)	
2. Average of top 5 performing models (out-of-sample)	0.661	0.491	0.521
	(0.085)	(0.050)	
3. Average of top 10 performing models (out-of-sample)	0.670	0.490	0.520
	(0.084)	(0.050)	
B. Standard TSTSLS			
4. Education only	0.628	0.351	0.345
	(0.076)	(0.051)	
5. Education + occupation	0.642	0.421	0.433
	(0.085)	(0.051)	
6. Education + occupation + province	0.676	0.430	0.455
	(0.074)	(0.054)	
7. Education + occupation + province + race	0.762	0.475	0.484
	(0.072)	(0.055)	
8. Education + occupation + province + race + interactions	0.452	0.388	0.587
	(0.075)	(0.075)	
Sample size	1,241	258	1,034

Source: PSLSD (1993) and NIDS (2008–2012)

Notes: Specifications 4–7 include only main effects, specification 8 includes all main effects and all pairwise interactions. Bootstrapped standard errors (reps 200) in parentheses

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 in-sample 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 the use of a machine learning approach to improve the standard two-sample two-stage method for estimating the intergenerational income elasticity in sub-optimal data conditions. Supervised machine learning algorithms minimize 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 such 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.

Acknowledgements We acknowledge financial support from the Italian Ministry of Education and Research, SIR Grant Project N. RBSI14KDMF.

Funding Open access funding provided by Università degli Studi Roma Tre within the CRUI-CARE Agreement.

Alternative algorithms to predict parental income

We compare the predictive ability of several methods that can be used to predict fathers' income. The methods considered belong both to the family of linear algorithms (OLS, Ridge regression, LASSO, relaxed LASSO, Elastic net) and to the family of non-linear algorithms (Boosted regression, Random forests).[18]

To assess the relative prediction ability of these methods we proceed in four steps:

1. We randomly split the sample into two subsets: the training set (80 % of the observations) and the hold-out set (20 %).
2. For methods that do not need any tuning (all OLS specifications): we estimate the models on the training sample and we store the estimated prediction functions.
3. For methods that need regularization/complexity choice: we select the most appropriate set of parameters by estimating the mean squared prediction error out-of-sample by 5-fold cross-validation.[19] We then use the tuning parameters that produce the smallest MSE to estimate the prediction model and we store the prediction function (both tasks only need the training set as defined in step 1).
4. We then estimate the prediction error in the hold-out sample for non-tuned and tuned learners stored in step 3 and 4.

The algorithms that need complexity choice are:

A. **Ridge regression**: The estimate was obtained using the *elasticregress* Stata module developed by Townsend (2017). We considering one hundred values of λ_s. The selected λ value obtained by 5-fold cross-validation is 0.649.
B. **LASSO**: similarly, we considered 100 λ_s calculated as in (A). The selected λ value obtained by 5-fold cross-validation is 0.128.
C. **Elastic net**: we consider 15 $\alpha_s \in [0, 1]$ and 100 λ_s calculated as in (A). The selected $\alpha - \lambda$ pair obtained by 5-fold cross-validation is 0.071–0.163.
D. **Relaxed LASSO**: in the first step, the LASSO is tuned considering 100 λ_s, as in (A). In the second step, the elastic net is tuned considering 15 α_s and 100 λ_s, as in (C). The selected λ value obtained by 5-fold cross-validation in the first step is 0.014; the selected $\alpha - \lambda$ pair obtained by 5-fold cross-validation in the second step is $0.071 - 0.051$.
E. **Boosted regression**: tuned across different shrinkage factors (10 values ranging from 0.1 to 1), number of iterations (10 values ranging from 10 to 100) and the fraction of training

[18] Note that Neural network was found to perform very poorly in this prediction task and is therefore not reported here.

[19] For each algorithm and each combination of tuning parameters, 5-fold cross-validation obtains unbiased estimate of the mean square error with the following steps: (1) randomly partition the sample in five folds, (2) exclude the first of the fold and train the algorithm on the remaining four fifths of observations, (3) use the estimated coefficients to predict the dependent variable in the excluded fold, (4) store the resulting mean squared error, (5) Repeat steps 2 to 4 for each of the remaining folds, (6) Compute the average prediction squared error in the 5 folds.

Table 3 In and out-of-sample R^2 for alternative algorithms in the US

	Hold-out sample			Training sample
	R-squared	95% C.I.		R-squared
OLS	23.38 %	9.49 %	37.27 %	34.41 %
Random forest	24.01 %	13.65 %	34.38 %	33.59 %
Boosted regression	24.10 %	11.24 %	36.95 %	33.69 %
Ridge regression	24.54 %	14.52 %	34.55 %	32.35 %
Lasso	25.59 %	14.64 %	36.55 %	31.71 %
Elastic net	25.60 %	15.28 %	35.92 %	31.26 %
Relaxed lasso	26.08 %	14.40 %	37.75 %	32.57%

Source: PSID (1982)

Notes: all algorithms are tuned by 5-fold cross validation as described in A-F. The out-of-sample R^2 is estimated on the hold-out sample. Normalized confidence intervals are obtained with 200 bootstrap iterations

observations used to fit an individual tree (10 values ranging from 0.1 to 1). The estimate was performed using the *boost* Stata module developed by Schonlau (2018). The selected combination of parameters obtained by 5-fold cross-validation is: shrinkage factor = 0.2; number of iterations = 50; fraction of training observations used to fit an individual tree = 0.9.

F. **Random forests**: tuned across the number of iterations (values considered: 10, 50, 100), the number of variables to randomly investigate (15 values from 1 to 29), the maximum depth (six values ranging from 0 to 50) and the minimum number of observations per leaf

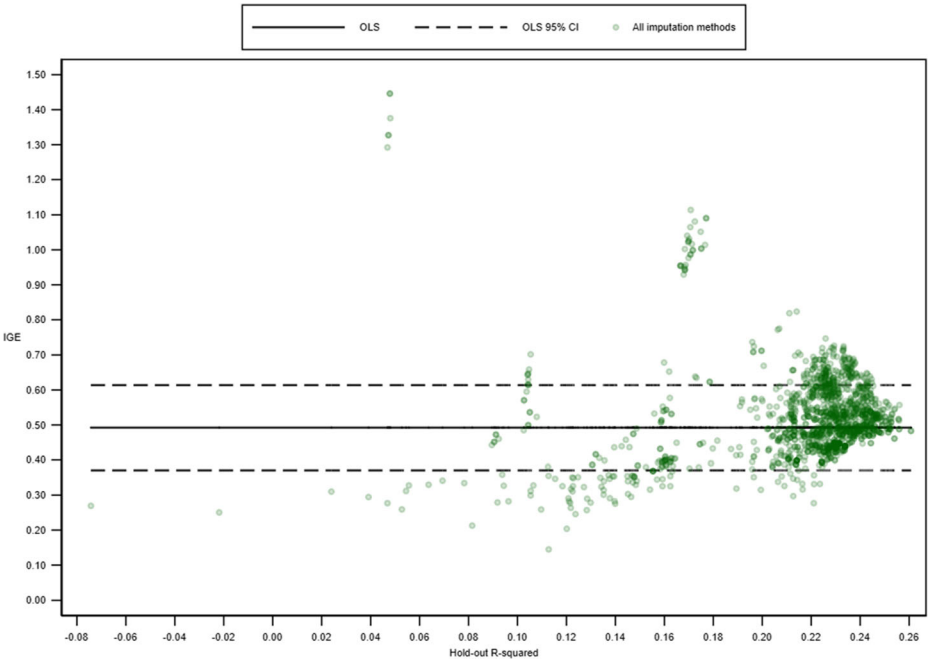

Fig. 7 First stage out-of-sample R^2 and IGE in the U.S. *Source:* PSID (1982) and PSID (2011)*Notes:* all algorithms are tuned by 5-fold cross validation as described in A-F. The out-of-sample R^2 is estimated on the hold-out sample. The solid horizontal line indicates the benchmark IGE estimated on longitudinal data (with the dashed lines displaying the 95% confidence interval)

Table 4 In and out-of-sample R^2 for alternative algorithms in South Africa

	Hold-out sample			Training sample
	R-squared	95% C.I.		R-squared
Random forest	48.80 %	39.26 %	58.33 %	52.68 %
OLS	48.97 %	38.61 %	59.33 %	52.02 %
Lasso	49.00 %	39.79 %	58.22 %	52.06 %
Ridge regression	49.06 %	40.22 %	57.90 %	53.91 %
Elastic net	49.11 %	39.27 %	58.95 %	49.99 %
Boosted regression	49.15 %	39.46 %	58.84 %	53.95 %
Relaxed lasso	49.37 %	39.97 %	58.77 %	52.89%

Source: PSLSD (1993)

Notes: all algorithms are tuned by 5-fold cross validation as described in A-F. The out-of-sample R^2 is estimated on the hold-out sample. Normalized confidence intervals are obtained with 200 bootstrap iterations

(values considered: 10, 30, 50). The estimate was performed using the *rforest* Stata module developed by Zou and Schonlau (2019). The selected combination of parameters obtained by 5-fold cross-validation is: number of iterations = 100; number of variables = 3; maximum depth = 50; minimum number of observations per leaf = 30.

Appendix Table 3 reports the performance of all algorithms for predicting parental income in the U.S. data. The first three columns contain the out-of-sample R^2 and the 95 % bootstrap confidence interval, while the last column reports the in-sample R^2. OLS refers to the best performing OLS specification among all possible combinations of regressors and includes as predictors: education, race, province and the pairwise interactions of education and race, and occupation and race. Note that the best OLS model outperforms all alternative algorithms *in sample*, but it is an overfitted model with a very large sampling variance and the lowest ability to predict out-of-sample. Moreover, the algorithms that performs well with non-linear data generating process (random forest and boosted regression) are outperformed by models that assume linearity: Ridge regression, LASSO and elastic net. The best performing model is relaxed LASSO which, as explained in the paper, uses two separate tuning parameters to perform variable selection and coefficient shrinkage.

It is important to note that confidence intervals are rather wide and overlap across the different models. This reflects some degree of uncertainty regarding the out-of-sample prediction error of the various algorithms due to relatively small sample sizes. While we opt for using the best-performing model in Appendix Table 3 for our empirical analysis, we do not claim this to be a definitive ranking of different ML algorithms in terms of their prediction ability.

Appendix Fig. 7 shows the relationship between out-of-sample R^2 and the IGE estimated in the second stage of the TSTSLS for all algorithms considered. It shows that the more accurate the out-of-sample prediction, the closer the estimated IGE is to the benchmark value obtained from the longitudinal data.

Appendix Table 4 displays the results for the South African sample of pseudo-fathers. In this case, all ML algorithms tend to work well in the hold-out sample, including the best OLS model.[20] Again, Relaxed LASSO is shown to be the preferred algorithm.

[20] OLS refers to the best performing OLS specification among all possible combinations of regressors and includes as predictors: education, race, province and the pairwise interactions of education and race, and occupation and race.

Additional tables

Table 5 Regularized IGEs in the U.S. by different first-stage sample size

	IGE	First-Stage R^2 (hold-out sample)	First-Stage R^2 (training sample)
1. Benchmark (OLS)	0.492		
	(0.062)		
Sample size	1,061		
Regularized TSTSLS using PSID (1981–1983) in the first-stage			
2. 'Best' model (hold-out sample)	0.501	0.292	0.375
	(0.081)	(0.036)	
3. Average of top 5 performing models (hold-out sample)	0.493	0.287	0.371
	(0.077)	(0.037)	
4. Average of top 10 performing models (hold-out sample)	0.490	0.285	0.368
	(0.076)	(0.037)	
Sample size	1,061	479	1,917

Source: PSID (1981-1983)

Notes: Bootstrapped standard errors (reps 200) in parentheses.

Table 6 IGE estimates: Regularized vs. standard TSTSLS including missing categories in the prediction set

	IGE	First stage R^2 (hold-out sample)	First stage R^2 (training sample)
1. Benchmark (OLS)	0.507		
	(0.054)		
A. Regularized TSTSLS			
2. 'Best' model (hold-out sample)	0.517	0.300	0.329
	(0.066)	(0.047)	
3. Average of top 5 performing models (hold-out sample)	0.508	0.299	0.328
	(0.066)	(0.047)	
4. Average of top 10 performing models (hold-out sample)	0.504	0.298	0.324
	(0.066)	(0.047)	
B. Standard TSTSLS			
5. Education only	0.762	0.116	0.193
	(0.087)	(0.042)	
6. Education + occupation	0.478	0.206	0.244
	(0.066)	(0.047)	
7. Education + occupation + industry	0.402	0.273	0.291
	(0.060)	(0.045)	
8. Education + occupation + industry + race	0.422	0.287	0.295
	(0.059)	(0.044)	
9. Education + occupation + industry + race + interactions	0.056	0.066	0.416
	(0.040)	(0.103)	
Sample size	1,199	384	1,535

Source: PSID longitudinal sample (1968-1992) and PSID (2011) for OLS benchmark; PSID (1982) and PSID (2011) for standard and regularized TSTSLS

Notes: Specifications 5-8 include only main effects, specification 9 includes all main effects and all pairwise interactions. Bootstrapped standard errors (reps 200) in parentheses

References

Aaronson, D., Mazumder, B.: Intergenerational economic mobility in the United States, 1940 to 2000. J. Hum. Resour. **43**(1), 139–172 (2008)

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

Athey, S., Imbens, G.W.: The state of applied econometrics: causality and policy evaluation. J. Econ. Perspect. **31**(2), 3–32 (2017)

Belloni, A., Chernozhukov, V., Hansen, C.: High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. **28**(2), 29–50 (2014)

Björklund, A., Jäntti, M.: Intergenerational income mobility in Sweden compared to the United States. Am. Econ. Rev. **87**(4), 1009–1018 (1997)

Björklund, A., Jäntti, M.: Intergenerational income mobility and the role of family background. In: Salverda, W., Nolan, B., Smeeding, T. (eds.) Handbook of Economic Inequality. Oxford University Press, Oxford (2009)

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

Blundell, J., Risa, E.: Income and family background: are we using the right models?Available at SSRN: https://ssrn.com/abstract=3269576 or https://doi.org/10.2139/ssrn.3269576 (2019)

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

Chen, W.-H., Ostrovsky, Y., Piraino, P.: Lifecycle variation, errors-in-variables bias and nonlinearities in intergenerational income transmission: new evidence from Canada. Labour Econ. **44**, 1–12 (2017)

Chetty, R., Hendren, N., Kline, P., Saez, E.: Where is the land of opportunity? The geography of intergenerational mobility in the United States. Quart. J. Econ. **129**(4), 1553–1623 (2014)

Clark, G.: The son also rises: surnames and the history of social mobility. Princeton University Press,Princeton (2014)

Corak, M.: Do poor children become poor adults? Lessons from a cross-country comparison of generational earnings mobility. Res. Econ. Inequality **13**(1), 143–188 (2006)

Corak, M.: Income inequality, equality of opportunity, and intergenerational mobility. J. Econ. Perspect. **27**(3), 79–102 (2013)

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**, 407–451 (2004)

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

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

Gong, H., Leigh, A., Meng, X.: Intergenerational income mobility in urban China. Rev. Income Wealth **58**(3), 481–503 (2012)

Haider, S., Solon, G.: Life-cycle variation in the association between current and lifetime earnings. Am. Econ. Rev. **96**(4), 1308–1320 (2006)

Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, Berlin (2009)

Hastie, T., Tibshirani, R., Tibshirani, R.J.: Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692(2017)

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

Jerrim, J., Choi, A., Simancas, R.: Two-Sample Two-Stage Least Squares (TSTSLS) estimates of earnings mobility: how consistent are they? Surv. Res. Methods **10**(2), 85–102 (2016)
McKenzie, D., Sansone, D.: Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria. J. Dev. Econ. **141**, 1–18.(2019)
Meinshausen, N.: Relaxed Lasso. Comput. Stat. Data Anal. **52**, 374–393 (2007)
Mullainathan, S., Spiess, J.: Machine learning: an applied econometric approach. J. Econ. Perspect. **31**(2), 87–106 (2017)
Narayan, A., Van der Weide, R., Cojocaru, A., Lakner, C., Redaelli, S., Mahler, D.G., Ramasubbaiah, R.G., Thewissen, S.: Fair Progress? In: Economic mobility across generations around the world. World Bank Group,Washington, DC (2018)
Nybom, M., Stuhler, J.: Heterogeneous income profiles and lifecycle bias in intergenerational mobility estimation. J. Hum. Resour. **51**(1), 239–268 (2016)
Olivetti, C., Paserman, D.: In the name of the son (and the Daughter): intergenerational mobility in the United States. Am. Econ. Rev. **105**(8), 1850–1940 (2015)
Piraino, P.: Intergenerational earnings mobility and equality of opportunity in South Africa. World Dev. **67**, 396–405 (2015)
Santavirta, T., Stuhler, J.: Name-based estimators of intergenerational mobility. Mimeo., Stockholm University (2020)
Schonlau, M.: BOOST: Stata module to perform boosted regression. Available at: https://ideas.repec.org/c/boc/bocode/s458541.html (2018)
Solon, G.: Intergenerational income mobility in the United States. Am. Econ. Rev. **82**(3), 393–408 (1992)
Solon, G.: Cross-country differences in intergenerational earnings mobility. J. Econ. Perspect. **16**(3), 59–66 (2002)
Townsend, W.: ELASTICREGRESS: Stata module to perform elastic net regression, lasso regression, ridge regression. Available at: https://ideas.repec.org/c/boc/bocode/s458397.html (2017)
Varian, H.: Big data: new tricks for econometrics. J. Econ. Perspect. **28**(2), 3–27 (2014)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B **67.2**, 301–320." (2005)
Zou, Y.R., Schonlau, M.: RFOREST: Stata module to implement Random Forest algorithm. Available at: https://ideas.repec.org/c/boc/bocode/s458614.html (2019)