The Journal of Economic Inequality (2025) 23:549–570
https://doi.org/10.1007/s10888-024-09643-8

RESEARCH

The two-sample two-stage least squares method to estimate the intergenerational earnings elasticity

Javier Cortes Orihuela[1] · Juan D. Díaz[1] · Pablo Gutiérrez Cubillos[1] · Pablo A. Troncoso[2]

Received: 1 November 2022 / Accepted: 30 July 2024 / Published online: 6 November 2024

Abstract
We show that the inconsistency of the Two-Sample Two-Stage Least Squares (TSTSLS) Intergenerational Earnings Elasticity (IGE) estimator is a two-way prediction problem involving the replication of (i) the variance of unobserved parental earnings, and (ii) the endogeneity of unobserved parental earnings in the equation of children's earnings. Concretely, we show that the TSTSLS estimator asymptotically recovers the OLS IGE when the first-stage R^2, i.e., the share of explained variance of parental earnings, equals the share of explained endogeneity of parental earnings in the child's earnings equation. This condition leads to two notable outcomes with respect to previous findings in the literature: (i) perfect prediction of parental earnings is a specific instance of our condition, indicating that consistency can be attained even when parental earnings are predicted imperfectly and (ii) exogenous instruments alone are insufficient to guarantee asymptotic equivalence between TSTSLS and OLS IGE estimates. Furthermore, our condition suggests that strong first-stage instruments might amplify TSTSLS bias if they are also strongly endogenous in the child's earnings equation. This last result provides a formal criterion for choosing first-stage predictors under the assumption that TSTSLS IGE estimates exhibit upward bias. Additionally, we theoretically study the biases of the two-sample stochastic multiple imputation and cell multiple imputation (MI) procedures, identifying conditions under which MI procedures outperform the traditional TSTSLS estimator. Finally, we validate our results through an empirical Monte Carlo exercise using administrative data from the Chilean formal private sector.

Keywords Intergenerational mobility · Linked administrative data · Two-sample two-stage least squares

✉ Pablo Gutiérrez Cubillos
pgutiecu@fen.uchile.cl

Javier Cortes Orihuela
jcorteso@fen.uchile.cl

Juan D. Díaz
juadiaz@fen.uchile.cl

Pablo A. Troncoso
patronco@central.uh.edu

[1] Department of Management Control and Information Systems, Faculty of Economics andBusiness, University of Chile, Santiago, Chile

[2] Department of Economics, University of Houston, Houston, TX 77004, U.S.

1 Introduction

The Intergenerational Earnings Elasticity (IGE) is a widely used metric for assessing intergenerational economic mobility (Corak 2013). It measures the increase in children's earnings associated with a one-percent increase in parental earnings, indicating the degree of economic persistence in a region, with a higher IGE implying greater socioeconomic persistence across generations.

This parameter is estimated using two methods, depending on data availability. In developed countries with access to linked administrative earnings data, it is estimated through ordinary least squares (OLS) regression.[1] In developing countries, where parental earnings data are often unavailable, researchers use the Two-Sample Two-Stage Least Squares (TSTSLS) estimator (Björklund and Jäntti 1997).[2] This method involves imputing missing parental earnings using available variables, such as education, age, or work experience.[3] Specifically, the TSTSLS uses two samples to estimate the IGE. The main sample includes data on parents' demographics and their children's earnings. The auxiliary sample consists of pseudo-parents, i.e., individuals from the same parents' cohort observed years earlier. This auxiliary sample contains data on earnings and demographic information, which are used in the first stage to estimate a Mincer equation. The estimated Mincer equation is then used to impute the earnings of the actual parents. In the second stage, children's earnings are regressed against imputed parental earnings. The resulting slope is referred to as the TSTSLS IGE.

The use of TSTSLS in the intergenerational context was originally inspired by Solon (1992), who used an instrumental variable (IV) estimator to provide an upper bound on the IGE. Since this work, the literature has extensively documented how TSTSLS IGE estimates differ from OLS IGE estimates. Specifically, Solon (1992), Björklund and Jäntti (1997), Nicoletti and Ermisch (2008), Jerrim et al. (2016), and Bloise et al. (2021) attribute these differences to two factors: (i) the instruments used in the Mincer equation to predict missing parental earnings also directly affect children's earnings (i.e., the instruments are endogenous), and (ii) the prediction of unobserved parental earnings is imperfect (i.e., the R^2 is not equal to one).

Building on this foundation, our paper makes three contributions. First, we establish a comprehensive framework for analyzing the TSTSLS IGE estimator, conceptualizing TSTSLS inconsistency as a two-way prediction problem involving both the estimation of the variance of unobserved parental earnings and the replication of the endogeneity of unobserved parental earnings in the child's earnings equation. Using our framework, we demonstrate that the TSTSLS IGE estimator converges to the OLS IGE when the share of explained variance of parental earnings equals the share of explained endogeneity of parental earnings in the child's earnings equation. We term this **the explained-variance-endogeneity equality condition**.

[1] Corak and Heisz (1999) is the pioneering work estimating the IGE through administrative records. More recently, Chetty et al. (2014) has led the resurgence of intergenerational mobility research based on administrative data. Additionally, Deutscher and Mazumder (2020) estimate intergenerational mobility for Australia, and Acciari et al. (2022) for Italy.

[2] Klevmarken (1982) develops a similar two-sample, two-stage procedure. The TSTSLS procedure is similar to the Two-Stage Instrumental Variable (TSIV) strategy (Angrist and Krueger 1992; Solon 1992). However, Inoue and Solon (2010) show that the TSTSLS estimator is asymptotically more efficient than the TSIV estimator.

[3] This method is widely used today. For example, Jerrim et al. (2016) presents a list of 30 papers using this method. Recently, Kenedi and Sirugue (2023) used it to study intergenerational mobility in France, and Jácome et al. (2023) employed an imputation procedure in the spirit of TSTSLS to study historical mobility in the U.S.

This framework yields two insights regarding conditions previously established in the literature. First, an exogenous instrument recovers the IGE only when parental earnings is an exogenous regressor in the children's earnings equation, which holds only if additional assumptions on the error term in the children's earnings equation are met. Second, perfect prediction of unobserved parental earnings (i.e., an R^2 of 1) implies our explained-variance-endogeneity equality condition, but the latter does not imply the former. Thus, perfect prediction of parental earnings is a particular case of our explained-variance-endogeneity equality condition.

Our second contribution provides a formal criterion for adding or removing first-stage predictors. Under the assumption that TSTSLS IGE exhibits upward bias with an initial set of commonly used predictors, the criterion is as follows: instruments should be included or removed if their inclusion or removal yields a percentage increase in the imputed endogeneity lower than the percentage increase in the first stage R^2. In other words, adding new variables improves TSTSLS bias if these variables enhance the variance of imputed parental earnings more than they enhance the replication of the endogenous component of parental earnings in the children's equation.

Our third contribution applies our framework to examine the asymptotic bias associated with estimating the IGE through multiple imputation procedures. We study stochastic multiple imputation (Cortes Orihuela et al. 2023) and cell multiple imputation (Jácome et al. 2023), finding conditions under which these methods outperform the traditional TSTSLS estimator. Using administrative data from the Chilean formal private sector, we perform an empirical Monte Carlo exercise (EMC) to illustrate our variable selection criterion and evaluate the performance of multiple imputation procedures relative to the traditional TSTSLS estimator.

The paper is organized as follows: Section 2 presents an overview of TSTSLS IGE estimation and reviews the existing literature. Section 3 presents a general framework to study TSTSLS consistency. Section 4 illustrates our framework by revisiting a classical TSTSLS IGE application from the literature and verifies our results through a theoretical Monte Carlo exercise. Section 5 studies the performance of two-sample multiple imputation procedures relative to the traditional TSTSLS estimator for IGE estimation. Section 6 shows empirical results of our framework through an empirical Monte Carlo exercise using administrative data from the Chilean private sector. Section 7 concludes.

2 The TSTSLS IGE estimator

In this section, we describe the IGE, the TSTSLS method, and discuss the conditions, as indicated in the literature, that induce convergence between the OLS IGE and TSTSLS IGE.

2.1 The intergenerational earnings elasticity

The intergenerational earnings elasticity is defined as:

$$\rho \equiv \frac{cov(y^c, y^p)}{var(y^p)}, \tag{1}$$

where y^c stands for the log permanent individual earnings of the child, y^p is the logarithm of his parents permanent earnings. This indicator is one of the most well-known intergenerational mobility measures, since it provides a statistical measure of the intergenerational association between the earnings of children and their parents (Corak 2013). A lower (higher)

IGE suggest a more (less) intergenerationally mobile society. Under ideal data conditions, the IGE is usually estimated through OLS regression.

2.2 Two-sample two-stage least squares estimator

In the absence of parental earnings, the Two-Sample Two-Stage Least Squares (TSTSLS) estimator has been widely used to measure intergenerational mobility. Specifically, the TSTSLS requires two samples: the primary sample includes data on the parents' demographics (such as education) and the earnings of their children, while the auxiliary sample consists of pseudo-parents. These pseudo-parents are earlier data from the parents' cohort with both earnings and demographic information. The TSTSLS method then proceeds in two steps using these two samples as a foundation. First, a Mincer equation is estimated by OLS using the auxiliary sample of pseudo-parents:

$$y_j^{pp} = \delta' z_j^{pp} + v_j, \tag{2}$$

where y_j^{pp} is the log-earnings of pseudo-parent j, z_j^{pp} is a vector of characteristics of the pseudo-parent, and v_j is the error term. After estimating δ', we impute y_i^p as $\hat{y}_i^p = \hat{\delta}' z_i{}^p$ where z_i^p is the vector of characteristics for the actual parent i. Then, the second stage of the TSTSLS estimator consists of regressing y_i^c on $\hat{y}_i^p$ by OLS. The estimated slope $\hat{\rho}^{\text{TSTSLS}}$ is the TSTSLS IGE.

2.3 Literature review on TSTSLS IGE estimation

We now discuss the development of the TSTSLS literature, which began with two seminal papers that initiated the study of the inconsistency of the TSTSLS estimator: Solon (1992) and Björklund and Jäntti (1997). One distinction between these works is that Solon (1992) focuses on the one-sample instrumental variable setting, while Björklund and Jäntti (1997) extend the analysis to the two-sample scenario. Despite this difference, their analyses share similarities as both assume that children's log-earnings depend on parental log-earnings and parental education. Consequently, they both conclude that the TSTSLS estimator consistently estimates ρ only under the condition that parental education does not influence children's status or when parental education and income are perfectly correlated. In their specific context, assuming that parental education does not affect children's earnings yields two key conditions: (i) children's earnings depend solely on parental earnings, ensuring that parental earnings are exogenous with respect to the error term in the children's equation, and (ii) parental education is an exogenous instrument. Importantly, in our paper, we demonstrate that if parental earnings are exogenous in the children's earnings equation, any relevant instrument will be exogenous, thereby inducing TSTSLS consistency. However, under more general assumptions, the use of an exogenous instrument does not ensure TSTSLS consistency. In this context, the scenarios explored in Solon (1992) and Björklund and Jäntti (1997) represent specific instances within our broader framework.

The next work that explores TSTSLS IGE inconsistency is Nicoletti and Ermisch (2008). In their framework, children's earnings depend on parental earnings as well as parents' occupational and educational characteristics. They conclude that TSTSLS is consistent when the first-stage R^2 equals one and provide a criterion for selecting first-stage predictors: chosen instruments should have minimal correlation with the error in the children's equation while maximizing multiple correlation with parental earnings. In our paper, we demonstrate that the perfect prediction of unobserved parental earnings induces our explained-variance-

endogeneity equality condition, but the latter does not imply the former. Hence, the condition proposed in Nicoletti and Ermisch (2008) is a specific case within our general framework. Moreover, through empirical and parametric Monte Carlo simulations, we not only affirm that the instrument selection criterion suggested by Nicoletti and Ermisch (2008) can be enhanced but also propose an alternative formal criterion based on the explained-variance-endogeneity equality condition.

The subsequent work in the literature is the study by Jerrim et al. (2016), which focuses on analyzing the bias of TSTSLS estimates with an empirical emphasis. They begin by summarizing the existing literature, asserting that TSTSLS is consistent when (i) the instrumental variables have no direct effect on children's earnings, and (ii) the R^2 of the first-stage equation equals 1. Subsequently, Jerrim et al. (2016) models children's earnings as a function of both actual parental earnings and imputed parental earnings. Notably, they observe that the inclusion of additional variables to enhance the R^2 of the first-stage prediction equation may simultaneously impact the effect of the predicted parental earnings. They emphasize that this issue could be particularly relevant when incorporating highly endogenous instruments, and it remains unclear whether any gains in prediction might be offset by this added endogeneity. Our paper extends the work of Jerrim et al. (2016), as we provide the exact condition under which TSTSLS IGE is consistent: whenever the R^2 equals the degree of endogeneity recovered through the imputed parental earnings, i.e., what we call the explained-variance-endogeneity equality condition.

Finally, Bloise et al. (2021), Cortes Orihuela et al. (2023), and Jácome et al. (2023) propose decompositions of the TSTSLS bias without assuming any particular data generating process for children's earnings. The works by Bloise et al. (2021) and Jácome et al. (2023) are similar, as they decompose TSTSLS bias into two sources: (i) a bias stemming from the imperfect prediction of parental earnings, and (ii) a bias that arises from the use of endogenous instruments to predict parental earnings. Similarly, Cortes Orihuela et al. (2023) decompose the TSTSLS IGE bias into two sources: a projection bias and a variance bias. They further propose a correction to the TSTSLS estimator which yields a lower bound for the OLS IGE and that eliminates the variance bias in TSTSLS estimates. Our work builds on these studies by clarifying the type of biases plaguing TSTSLS estimates, as we frame TSTSLS inconsistency as a two-way problem involving: (i) the correct prediction of parental earnings variance, and (ii) the replication of the endogeneity of parental earnings in the children's earnings equation. We further show conditions involving these two prediction biases under which TSTSLS IGE is consistent.

3 A general framework to understand TSTSLS consistency

In this section, we develop a general framework to examine TSTSLS consistency, establishing conditions for consistency that depend on the exogeneity of parental earnings in the children's earnings equation. Then, we explore how, unless parental earnings are exogenous in the children's earnings equation, the utilization of exogenous instruments falls short in achieving TSTSLS consistency. Building on our conditions for consistency, we then propose a formal criterion for instrument selection grounded in plausible assumptions.

3.1 Revisiting consistency of the IGE using TSTSLS

We work under the assumption that the sample of pseudo-parents is asymptotically representative of the distribution of actual parents. We also follow (Jerrim et al. 2016) by assuming that the covariance between imputed and actual parental log-earnings is asymptotically equal

to the variance of imputed parental log-earnings. Then, the IGE is estimated in most empirical applications using the following Galtonian regression model:

$$y_i^c = \alpha + \beta y_i^p + \psi_i, \tag{3}$$

where α is the intercept, β would be a causal effect absent any endogeneity of y^p on y^c, and ψ_i is the stochastic error term. We can apply the OLS orthogonal decomposition to Eq. 3:

$$y_i^c = \hat{\alpha} + \hat{\rho} y_i^p + e_i, \tag{4}$$

where $\hat{\alpha}$ is the OLS estimator of α, $\hat{\rho}$ is the OLS estimator of β, and e_i is the residual term such that $cov(y_i^p, e_i) = 0 \; \forall i$. We begin by noting that $\hat{\rho}$ is consistent to estimate the true IGE parameter

$$\text{plim}\, \hat{\rho} = \frac{cov(y_i^c, y_i^p)}{var(y_i^p)} = \rho. \tag{5}$$

by the Weak Law of Large Numbers and the definition of the IGE stated in Eq. 1. This result is independent of any assumptions about the structure of the error term ψ_i. However, as we will show, unlike OLS consistency, TSTSLS consistency with respect to ρ depends on the structure of this error term.[4]

Case 1: Parental earnings in the IGE equation are exogenous

In Eq. 3, assume that $E(\psi_i | y_i^p) = 0$. That is, there are no variables that affect children's earnings which are also correlated with parental earnings. In this context, β is equal to ρ, meaning, the causal effect of y^p on y^c is equal to the IGE. Furthermore, assume that we estimate the IGE parameter through an imputation of parental earnings $\hat{y}_i^p$ which is built from a TSTSLS approach.[5] Then, the TSTSLS estimator converges to:

$$\text{plim}\, \hat{\rho}^{TSTSLS} = \frac{cov(y_i^c, \hat{y}_i^p)}{var(\hat{y}_i^p)} = \frac{\beta \cdot cov(y_i^p, \hat{y}_i^p) + cov(\psi_i, \hat{y}_i^p)}{var(\hat{y}_i^p)} = \beta = \rho. \tag{6}$$

This is because (i) $cov(\hat{y}_i^p, y_i^p) = var(\hat{y}_i^p)$, by our assumption in the spirit of Jerrim et al. (2016), and (ii) $cov(\hat{y}_i^p, \psi_i) = 0$ because of $E(\psi_i | y_i^p) = 0$. That is, when parental earnings is an exogenous regressor both TSTSLS and OLS converge to the same parameter for any instrument used in the first stage.

Case 2: Parental earnings in the IGE equation are endogenous

Assume the general case in which the data generating process of y_i^c is given by:

$$y_i^c = \alpha + \beta y_i^p + \varepsilon_i, \tag{7}$$

where $\varepsilon_i = \varphi_i + \psi_i$. Assume that $E(\varphi_i | y_i^p) \neq 0$ and $E(\psi_i | y_i^p) = 0$. That is, the error term ε_i is composed by an endogenous component φ_i, and an exogenous component ψ_i. Then, the OLS estimate of the IGE asymptotically converges to:

$$\text{plim}\, \hat{\rho} = \frac{cov(y_i^c, y_i^p)}{var(y_i^p)} = \beta + \frac{cov(y_i^p, \varphi_i)}{var(y_i^p)} = \rho. \tag{8}$$

As we can see, $\hat{\rho}$, which is our estimation of interest, does not converge to β which is the causal effect because $cov(y_i^p, \varphi_i) \neq 0$. Indeed, we can interpret $cov(y_i^p, \varphi_i)$ as a measure of

[4] For notation convenience we set $\text{plim}\, \hat{\beta} = \text{plim}\, \hat{\rho} = \rho$.

[5] We assume that the instruments used to impute y_i^p are a subset of the variables that participate in the data generating process for y_i^p. This assumption is relevant, as using other instruments such as collider variables, would alter the results that follow due to endogenous selection bias.

the endogeneity of parental earnings present in Eq. 3. In this context, the TSTSLS estimator converges to:

$$\text{plim}\ \hat{\rho}^{TSTSLS} = \frac{cov(y_i^c, \hat{y}_i^p)}{var(\hat{y}_i^p)} = \frac{\beta \cdot cov(y_i^p, \hat{y}_i^p) + cov(\varphi_i, \hat{y}_i^p)}{var(\hat{y}_i^p)} = \beta + \frac{cov(\varphi_i, \hat{y}_i^p)}{var(\hat{y}_i^p)}, \quad (9)$$

$$= \rho + \left[\frac{cov(\hat{y}_i^p, \varphi_i)}{var(\hat{y}_i^p)} - \frac{cov(y_i^p, \varphi_i)}{var(y_i^p)} \right], \quad (10)$$

where we use Eq. 8. Using Eq. 10 we can state the following proposition:

Proposition 1 *The TSTSLS IGE estimator is consistent when* $\frac{cov(\hat{y}_i^p, \varphi_i)}{cov(y_i^p, \varphi_i)} = R^2$*, where* R^2 *is the goodness-of-fit that comes from the first stage regression between pseudo-parents' earnings and demographic variables*

Proof From Eq. 10, the asymptotic bias in the TSTSLS IGE estimator is given by:

$$\frac{cov(\hat{y}_i^p, \varphi_i)}{var(\hat{y}_i^p)} - \frac{cov(y_i^p, \varphi_i)}{var(y_i^p)} = \text{plim}\ \hat{\rho}^{TSTSLS} - \text{plim}\ \hat{\rho}, \quad (11)$$

which in turn yields the following condition for TSTSLS consistency:

$$\frac{cov(\hat{y}_i^p, \varphi_i)}{var(\hat{y}_i^p)} = \frac{cov(y_i^p, \varphi_i)}{var(y_i^p)} \iff \frac{cov(\hat{y}_i^p, \varphi_i)}{cov(y_i^p, \varphi_i)} = \frac{var(\hat{y}_i^p)}{var(y_i^p)} = R^2. \quad (12)$$

□

That is, the TSTSLS IGE estimator is consistent when the share of explained variance of parental earnings and the proportion of explained endogeneity of parental earnings are equal. We emphasize that this condition, which we call the explained-variance-endogeneity equality condition, has not been formally discussed in the literature.[6] To further understand this condition, note that $\hat{\rho}^{TSTSLS}$ is equal to ρ when the ratio of $cov(\varphi_i, y_i^p)$ and $var(y_i^p)$ can be replicated by the ratio of $cov(\varphi_i, \hat{y}_i^p)$ and $var(\hat{y}_i^p)$. That is, if $var(y_i^p)$ is different than $var(\hat{y}_i^p)$, then $cov(\varphi_i, y_i^p)$ should be different than $cov(\varphi_i, \hat{y}_i^p)$ in the same proportion. Recall that the proportional difference between $var(\hat{y}_i^p)$ and $var(y_i^p)$ is the first-stage R^2, and by definition, the proportional difference between $cov(\varphi_i, \hat{y}_i^p)$ and $cov(\varphi_i, y_i^p)$ is $\frac{cov(\varphi_i, \hat{y}_i^p)}{cov(\varphi_i, y_i^p)}$. This means that if the variance of the predicted parental earnings is low, then to recover ρ via $\hat{\rho}^{TSTSLS}$, the explained endogeneity by the predicted parental earnings should be low.

In this context, TSTSLS inconsistency is a two-way prediction problem involving: (i) the prediction of parental earnings variance, represented by R^2 in Eq. 12, and (ii) the prediction of the endogeneity of parental earnings with respect to the equation of children's earnings, given by $\frac{cov(\hat{y}i^p, \varphi i)}{cov(y_i^p, \varphi_i)}$ in Eq. 12.[7] Notably, it can be seen that an R^2 of 1 implies that the equality stated in Eq. 12 holds. However, the latter does not imply the former, and so the usual $R^2 = 1$ condition established in the literature is a particular case of our general statement. Finally, we note that this condition suggests that the TSTSLS IGE can be consistent even when $R^2 < 1$.

[6] As stated previously, Jerrim et al. (2016) hint at a trade-off in the endogeneity and prediction gains from adding or removing first-stage instruments. Nonetheless, their work does not formalize this intuition.

[7] Recall that we have assumed that the covariance between imputed and actual parental log-earnings is asymptotically equal to the variance of imputed parental log-earnings, and so the R^2 of the first-stage is equivalent to the R^2 of imputed parents in the main sample. However, this might not hold in actual two-sample applications. For example, overfitting might yield a high R^2 in the first-stage but a low out-of-sample R^2. Similarly, samples might be heterogeneous.

3.2 Exogenous instruments and the IGE

As reviewed in Section 2.3, the literature has suggested that the differences between the TSTSLS IGE and the OLS IGE can be reconciled through the use of an exogenous instrument in the first stage of the TSTSLS estimation. For example, Acciari et al. (2022) recently use linked administrative data for Italy and find that the TSTSLS IGE estimate is more than double the OLS IGE estimate. They suggest that the key challenge in TSTSLS estimation is "finding a valid instrument, one that predicts parental income but remains orthogonal to child income" (Acciari et al. 2022, p. 137).

We can use our framework to study the effect of including such an instrument on the TSTSLS IGE bias. Concretely, assume that we have an instrument z that is used in the first stage of a TSTSLS IGE estimation. Further, assume that $E(\varepsilon_i|z_i) = 0$; that is, the instrument is exogenous with respect to the children's earnings equation. Let $\hat{\hat{y}}_i^p$ denote the first stage prediction using this instrument, we know that the TSTSLS IGE using an exogenous instrument converges to:

$$\text{plim}\,\hat{\hat{\rho}}^{TSTSLS} = \beta + \frac{cov(\hat{\hat{y}}_i^p, \varphi_i)}{var(\hat{\hat{y}}_i^p)} = \beta. \tag{13}$$

This is because $cov(\hat{\hat{y}}_i^p, \varphi_i) = 0$ given that $E(\varepsilon_i|z_i) = 0$. That is, $\hat{\hat{\rho}}^{TSTSLS}$ asymptotically converges to β:

$$\beta - \rho = -\frac{cov(y_i^p, \varphi_i)}{var(y_i^p)}. \tag{14}$$

This is because the use of an exogenous instrument, by construction, fully neglects the replication of the endogeneity of parental earnings in the children's equation. Specifically, an exogenous but relevant instrument yields $cov(\hat{\hat{y}}_i^p, \varphi_i) = 0$ and $R^2 > 0$, meaning that the share of explained endogeneity is null, while the share of explained variance is positive. Moreover, under the plausible assumption that $cov(y_i^p, \varphi_i) > 0$, the use of an exogenous instrument results in a lower bound for the IGE.

3.3 A formal criterion for TSTSLS instrument selection

The literature has documented that TSTSLS IGE estimates usually exhibit upward bias with respect to OLS estimates (Jerrim et al. 2016; Bloise et al. 2021; Zimmerman 1992; Solon 1992; Cortes Orihuela et al. 2023). From our results, it follows that TSTSLS estimates will overstate the true IGE when the share of explained endogeneity is greater than the share of explained variance. This suggests that we can decrease TSTSLS IGE bias by adding variables which are predictive for parental earnings, but not too endogenous with respect to the children's equation. Alternatively, we can improve the estimation of the IGE using TSTSLS by removing variables that are highly endogenous and have low predictive value for parental earnings.

Given this intuition, we use our framework to develop a formal criterion for the addition and/or removal of first-stage predictors in a TSTSLS setting. Concretely, assume that we add an additional variable, S which is used, alongside the previous variables, to generate a new prediction $\hat{y}'^p$. First, we can see that the R^2 in the first stage increases to $R'^2 \equiv R^2 + \Delta R^2$ with $\Delta R^2 \geq 0$. However, this additional predictor also changes the endogeneity of the imputed parental earnings. Define $cov(\hat{y}_i'^p, \varphi_i) \equiv cov(\hat{y}_i^p, \varphi_i) + \Delta cov(\hat{y}_i^p, \varphi_i)$, where, the sign of

$\Delta cov(\hat{y}_i^p, \varphi_i)$ is unknown. Under the plausible assumption that the endogeneity of parental earnings is positive, i.e., $cov(y_i^p, \varphi_i) > 0$, the following proposition defines a criterion on whether an additional predictor in the TSTSLS improves the estimation of the IGE.

Proposition 2 *Under the assumption that* $cov(y_i^p, \varphi_i) > 0$*, adding new variables improves the TSTSLS IGE asymptotic bias:*

i) When the TSTSLS IGE bias is positive, and $\frac{\Delta cov(\hat{y}_i^p, \varphi_i)}{cov(\hat{y}_i^p, \varphi_i)} < \frac{\Delta R^2}{R^2}$

ii) When the TSTSLS IGE bias is negative, and $\frac{\Delta cov(\hat{y}_i^p, \varphi_i)}{cov(\hat{y}_i^p, \varphi_i)} > \frac{\Delta R^2}{R^2}$

The proof of Proposition 2 can be found in the Online appendix. As we can see, the effect of adding a new variable in the TSTSLS IGE bias depends on the relationship between $\frac{\Delta cov(\hat{y}_i^p, \varphi_i)}{cov(\hat{y}_i^p, \varphi_i)}$ and $\frac{\Delta R^2}{R^2}$, i.e., whether the difference between the percent increase in the endogeneity captured by the imputation is greater or lower than the percent increase in the R^2 of the first stage. For example, assume that the TSTSLS IGE bias is positive, which is consistent with the literature presumptions and findings (Jerrim et al. 2016; Bloise et al. 2021; Zimmerman 1992; Solon 1992; Cortes Orihuela et al. 2023), we have that adding new variables reduces the TSTSLS bias if this additional variable yields a relative improvement in the parental prediction variance which is greater than the relative improvement in the replication of the endogeneity of parental earnings in the children's equation. From this analysis, we can also see that an increase in the R^2 of the first stage in the TSTSLS does not guarantee that the TSTSLS IGE will become closer to the OLS IGE. This is because the positive effect of an increase in the R^2 could be offset by an increase in the endogeneity in the children equation captured by this new prediction.[8]

4 Revisiting the classical TSTSLS application: exemplifying our framework

In this section, we exemplify our framework by making explicit assumptions about the data generating process for parent's and children's earnings.

4.1 The classical TSTSLS application

First, we briefly examine the model discussed in Zimmerman (1992); Solon (1992) and Björklund and Jäntti (1997). That is, we assume that the logarithm of children's earnings is a linear function of parental log-earnings and parental education ($Educ^p$):

$$y_i^c = \alpha + \beta \cdot y_i^p + \gamma \cdot Educ_i^p + \epsilon_i.$$

We further assume that parental earnings are a linear function of parental education:

$$y_i^p = \delta_0 + \delta_1 \cdot Educ_i^p + \eta_i. \tag{15}$$

We know that the OLS estimator for ρ converges to the true IGE:

$$\text{plim}\, \hat{\rho} = \beta + \gamma \cdot \frac{cov(y^p, Educ^p)}{var(y^p)} = \rho \tag{16}$$

[8] In the online appendix we extend our theoretical results to the presence of classical measurement error.

This formulation follows from the usual omitted variable bias formula in OLS estimation. Next, the TSTSLS IGE estimator (using education in the first stage) for ρ converges to:

$$\text{plim}\,\hat{\rho}^{\text{TSTSLS}} = \beta + \gamma \cdot \frac{cov(\hat{y}^p, Educ^p)}{var(\hat{y}^p)} \tag{17}$$

$$= \beta + \gamma \cdot \frac{cov(y^p, Educ^p)}{var(y^p)} \cdot \left[1 + \frac{1 - R^2}{R^2}\right], \tag{18}$$

Equation 18 suggests that, in this limited context, the conditions widely established in the literature hold: both methods converge to the same parameters if either (i) parental education has no direct effect on children's earnings ($\gamma = 0$) or (ii) we are able to perfectly predict the missing parental earnings ($R^2 = 1$). Moreover, notice that in this scenario our explained-variance-endogeneity equality condition reduces to $R^2 = \frac{var(\hat{y}^p)}{var(y^p)} = \frac{cov(\hat{y}^p, Educ^p)}{cov(y^p, Educ^p)} = 1$, where the last equality comes from the orthogonal decomposition, which is implied by the assumption in the spirit of Jerrim et al. (2016). Thus, our condition leads to the usual $R^2 = 1$ proviso in this simple model. Next, we expand this simple example by adding further endogeneity to the classical setting.

4.2 Adding more endogeneity to the classical example

We describe the asymptotic bias of TSTSLS IGE under additional endogeneity. Concretely, we first assume that the logarithm of children's earnings is generated by a linear function of parental log-earnings, parental education, and parental skill:

$$y_i^c = \alpha + \beta \cdot y_i^p + \gamma \cdot Educ_i^p + \theta \cdot Skill_i^p + \epsilon_i.$$

In an empirical TSTSLS context, $Educ$ and $Skill$ are usually observable and unobservable variables, respectively. Because of this, the researcher estimates the following function to impute the missing parental earnings,:

$$y_i^p = \delta_0 + \delta_1 \cdot Educ_i^p + \eta_i.$$

Finally, we assume that the education and skill of the parent are not independent. Instead, they are related through the following process:

$$Skill_i^p = \omega_0 + \omega_1 \cdot Educ_i^p + \xi_i.$$

Under these assumptions, the probability limit of the OLS IGE estimator follows the traditional omitted variable formula:

$$\text{plim}\,\hat{\rho} = \beta + \gamma \cdot \frac{cov(y^p, Educ^p)}{var(y^p)} + \theta \cdot \frac{cov(y^p, Skill^p)}{var(y^p)}. \tag{19}$$

Meanwhile, the TSTSLS IGE estimator converges to:

$$\text{plim}\,\hat{\rho}^{\text{TSTSLS}} = \beta + \gamma \cdot \frac{cov(\hat{y}^p, Educ^p)}{var(\hat{y}^p)} + \theta \cdot \frac{cov(\hat{y}^p, Skill^p)}{var(\hat{y}^p)}. \tag{20}$$

$$= \beta + \gamma \cdot \frac{cov(y^p, Educ^p)}{var(y^p)}\left[1 + \frac{1 - R^2}{R^2}\right] + \theta \cdot \omega_1 \frac{cov(y^p, Educ^p)}{var(y^p)}\left[1 + \frac{1 - R^2}{R^2}\right]. \tag{21}$$

Now, suppose that $\gamma = 0$ and $\omega_1 \neq 0$; that is, there is no direct effect of parental education on the child's earnings and that skill and education of the parent are correlated. From Eqs. 19

and 21, it can be seen that the estimation methods converge to different parameters. Hence, this analytical exercise illustrates that the use of exogenous instruments is not sufficient to ensure asymptotic equality between both estimators in a more general context. Concretely, by equating Eqs. 19 and 21 we obtain our explained-variance-endogeneity equality condition:

$$R^2 = \frac{var(\hat{y}^p)}{var(y^p)} = \frac{\gamma \cdot cov(\hat{y}^p, Educ^p) + \theta \cdot cov(\hat{y}^p, Skill^p)}{\gamma \cdot cov(y^p, Educ^p) + \theta \cdot cov(y^p, Skill^p)}. \tag{22}$$

That is, the TSTSLS estimator is consistent when the share of the explained variance in parental earnings equals the share of explained endogeneity of parental earnings in the children's earnings equation. Note that, unlike the classical scenario, the ratio of covariances is not 1 because *Skill* is unobservable, and so it is not included in the first-stage equation. This implies that the TSTSLS IGE estimator can be consistent even if $R < 1$.[9] Next, we perform a series of Monte Carlo simulations to assess our criterion for TSTSLS consistency.

4.3 Monte Carlo simulations

To show how the share of explained variance of parental earnings and the share of explained endogeneity of parental earnings in the TSTSLS procedure drives the differences between the OLS IGE and the TSTSLS IGE, we perform Monte Carlo simulations. In our exercise, we define two cases for children's earnings:

Case 1: $\gamma \neq 0$ and $\theta \neq 0$: Parental education has a direct and indirect effect on children's earnings

$$y_i^{c1} = 5 + 0.3 \cdot y^p + 0.3 \cdot Educ_i^p + 0.3 \cdot Skill_i^p + \epsilon_i. \tag{23}$$

Case 2: $\gamma = 0$ and $\theta \neq 0$: Parental education has only an indirect effect on children's earnings

$$y_i^{c2} = 5 + 0.3 \cdot y^p + 0.3 \cdot Skill_i^p + \epsilon_i. \tag{24}$$

That is, in the first case we assume that parental education has a direct effect on children's earnings, while in the second we assume that it affects children's earnings indirectly though y^p and $Skill^p$. We further assume that the data generating process for parental earnings depends on an instrument z, $Educ^p$, and $Skill^p$:

$$y_i^p = 10 + 4 \cdot z + 0.5 \cdot Educ_i^p + 0.5 \cdot Skill_i^p. \tag{25}$$

We consider that the covariance between $Educ_i^p$ and $Skill^p$ is positive.

$$Skill_i^p = 10 + 3 \cdot x_1 + 0.3 \cdot Educ_i^p + \xi_i. \tag{26}$$

Finally, we assume that $z \sim N(0, 1), x_1 \sim N(0, 1), \xi_i \sim N(0, 1)$ and $Educ_i^p \sim N(10, 9)$, which are all independent of each other. With this structure, the IGE in case 1 is $\rho^1 = 0.46955$, while in case 2 is $\rho^2 = 0.387$.

We investigate the relationship between $\hat{\rho}^{TSTSLS}$, the R^2 of the first stage and the share of explained endogeneity of parental earnings. To do so, we use the following predictor in the first stage:

$$f_i = a_1 \cdot 4 \cdot z_i + a_2 \cdot Educ_i^p + a_3 \cdot Skill_i^p, \tag{27}$$

[9] Notice that the perfect prediction of the missing parental earnings implies that $\omega_1 \frac{cov(y^p, Educ^p)}{var(y^p)} = \frac{cov(y^p, Skill^p)}{var(y^p)}$ which implies that $\frac{var(\hat{y}^p)}{var(y^p)} = \frac{\gamma \cdot cov(\hat{y}^p, Educ^p) + \theta \cdot cov(\hat{y}^p, Skill^p)}{\gamma cov(y^p, Educ^p) + \theta \cdot cov(y^p, Skill^p)} = 1$.

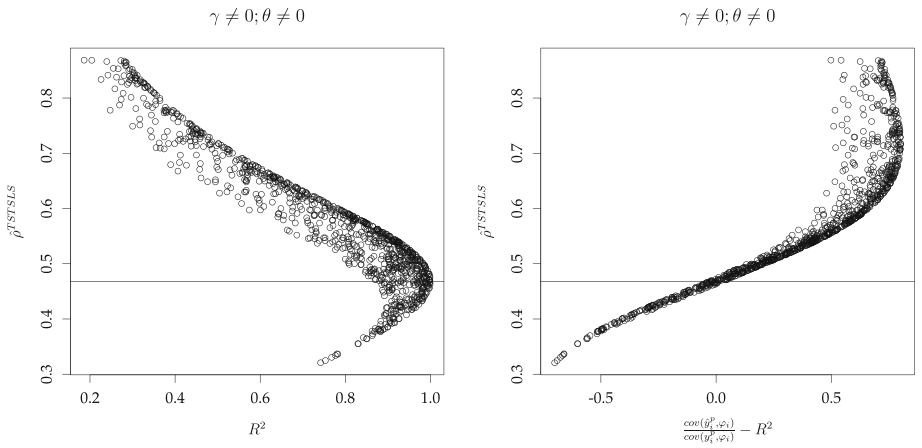

Fig. 1 TSTSLS IGE estimates vs R^2 in the first stage when $\gamma \neq 0$ and $\theta \neq 0$ (Case 1)†. †*Note*: The left panel shows the estimate of the TSTSLS IGE $\hat{\rho}^{TSTSLS}$ in the y-axis when we use f_i as a predictor of parental earnings and the R^2 of the first stage in the x-axis. The right panel shows the estimate of the TSTSLS IGE $\hat{\rho}^{TSTSLS}$ in the y-axis and the difference between the share of explained endogeneity and the R^2 of the first stage in the x-axis. The vertical black line denotes the OLS IGE estimate

where $a_i \sim \mathcal{U}(0, 1)$ for $i = 1, 2, 3$. We repeat this exercise 1,000 times. Here f_i is a distorted measure of parental earnings. In each iteration, we have a different $\hat{\rho}^{TSTSLS}$ with a different R^2 in the first stage of the TSTSLS procedure and a different share of explained endogeneity. In the following figures, the subfigures located in the left show the relationship between the IGE estimated using TSTSLS and the R^2 of the first stage, while the subfigures placed in the right depict the relationship between the IGE estimated using TSTSLS and the difference between the share of explained endogeneity and the R^2. Horizontal lines depict the true IGE. Figure 1 plots those relationships for the case when $\gamma \neq 0$ and $\theta \neq 0$ and Fig. 2 represents those relationships for the case when $\gamma = 0$ and $\theta \neq 0$.

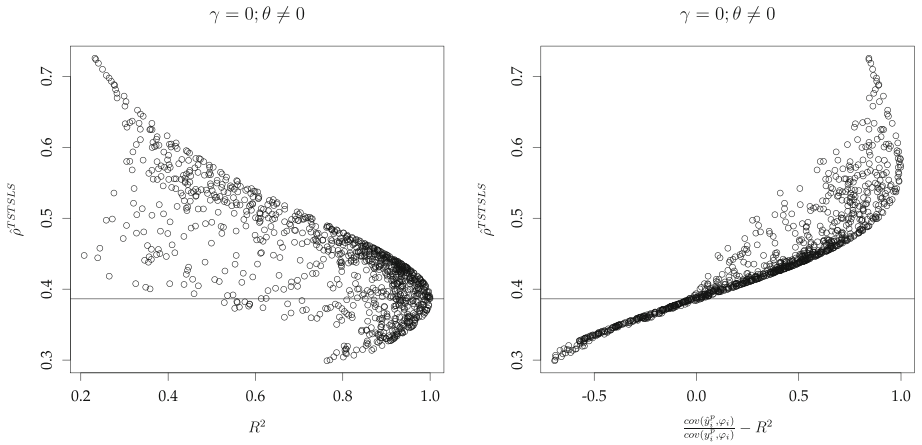

Fig. 2 TSTSLS IGE estimates vs R^2 in the first stage when $\gamma = 0$ and $\theta \neq 0$ (Case 2)†. †*Note*: The left panel shows the estimate of the TSTSLS IGE $\hat{\rho}^{TSTSLS}$ in the y-axis when we use f_i as a predictor of parental earnings and the R^2 of the first stage in the x-axis. The right panel shows the estimate of the TSTSLS IGE $\hat{\rho}^{TSTSLS}$ in the y-axis and the difference between the share of explained endogeneity and the R^2 of the first stage in the x-axis. The vertical black line denotes the OLS IGE estimate

First, the left panels of Figs. 1 and 2 suggest that as the R^2 of the first stage gets closer to 1, the TSTSLS method gets closer to the true IGE. However, an increase in the R^2 of the first stage in the TSTSLS does not guarantee TSTSLS IGE consistency. This is particularly evident in case 2 depicted by Fig. 2, where it can be observed that in some simulations a very low R^2 can still get very close to the true IGE. These scenarios are explained by the right panels, which suggest that the TSTSLS IGE equals the true IGE when the difference between the share of explained variance of parental earnings and the share of explained endogeneity of parental earnings is null. That is, TSTSLS inconsistency is not uniquely driven by the prediction of parental earnings, but also by the imputed endogeneity of parental earnings.

We also highlight that in case 1 it is less likely to obtain a TSTSLS that is very close to the true IGE with a low R^2. This is because in case 1 each variable that enters into the children's earnings equation is also part of the data generating process of parental earnings. Thus, prediction gains in the first stage generated by the variables $Educ$ and $Skill$ are more likely to be offset by their added endogeneity with respect to the children's equation. Whereas in case 2, $Educ$ is not as endogenous, as it has no direct effect over children's earnings. This implies that our explained-variance-endogeneity equality condition can hold even with a low R^2.[10]

5 Two-sample multiple imputation intergenerational earnings elasticity

In this section we use our framework to study the asymptotic bias of using multiple imputation (MI) procedures to predict unobserved parental earnings. Multiple imputation techniques have been widely used in statistical research to impute missing data (Rubin 2004). Recently, multiple imputation techniques have garnered attention within the literature of intergenerational mobility in contexts of data scarcity where parental earnings are not observed (Jácome et al. 2023) but also due to their potential usefulness in improving TSTSLS estimates (Cortes Orihuela et al. 2023). This technique is based on building M imputed datasets. Then, the parameter of interest is estimated M times using each imputed dataset. Finally, the multiple imputation IGE consists of averaging the estimated parameters among all imputed datasets. Concretely, we study two methods of multiple imputation used in intergenerational mobility: (i) the stochastic imputation method (Cortes Orihuela et al. 2023) and (ii) the grid multiple imputation method (Jácome et al. 2023).

5.1 Stochastic multiple imputation

In the context of TSTSLS with missing parental earnings data, it is possible to use stochastic imputation (Rubin 2004; Cortes Orihuela et al. 2023), i.e., the use of a regression imputation combined with a stochastic error drawn from a normal distribution. Formally, first define the $m \in \{1, ..., M\}$ parental imputation as:

$$\tilde{\hat{y}}^{p}_{i,m} = \hat{y}^{p}_{i,m} + \zeta_{i,m}, \tag{28}$$

where $\tilde{\hat{y}}^{p}_{i,m}$ is the m stochastic imputation, $\hat{y}^{p}_{i,m}$ is the first stage parental imputation using TSTSLS, $\zeta_{i,m} \sim \mathcal{N}\left(0, \sigma_v^2\right)$ is an error term drawn from a normal distribution with mean 0

[10] In the online appendix we deliver more details regarding the estimates from our Montecarlo simulations. We also explore the performance of exogenous instruments, showing that they fail to recover the true IGE.

and variance σ_v^2 which is the variance of the first stage parental earnings imputation equation residual v_j. Thus, we can estimate the IGE using each $\tilde{\hat{y}}_{i,m}^p$ as parental earnings and then compute the average of those estimates. For each m we call the IGE as $\tilde{\hat{\rho}}_m^{TSTSLS}$. Then, we have that the TSTSLS stochastic multiple imputation (SMI) IGE is given by

$$\hat{\rho}^{TSTSLS-SMI} \equiv \frac{1}{M}\sum_{m=1}^{M}\tilde{\hat{\rho}}_m^{TSTSLS}. \tag{29}$$

We can show conditions under which the SMI IGE can improve the estimation of the IGE relative to the traditional TSTSLS.

Proposition 3 *Stochastic multiple imputation improves the estimation of the IGE compared to baseline TSTSLS if the R^2 of the first stage holds:*

$$\frac{-\left[\frac{cov(\hat{y}_i^p,\varphi_i)}{var(\hat{y}_i^p)}-\frac{cov(y_i^p,\varphi_i)}{var(y_i^p)}\right]+\rho}{\left[\frac{cov(\hat{y}_i^p,\varphi_i)}{var(\hat{y}_i^p)}-\frac{cov(y_i^p,\varphi_i)}{var(y_i^p)}\right]+\rho} < R^2. \tag{30}$$

The proof of Proposition 3 can be found in the Online appendix. Thus, this multiple imputation procedure only improves the estimation of the TSTSLS IGE when the R^2 is sufficiently large. This improvement is attributed to the scaling of the TSTSLS IGE by the R^2 during multiple imputation. This is explained by the fact that the stochastic imputation procedure recovers the variance of the true parental earnings using a noise $\zeta_{i,m}$ that is uncorrelated with y^c. Thus, when the R^2 is too low, the estimated parameter under multiple imputation is low. We further note that the TSTSLS SMI IGE can serve as a lower bound for the OLS:

Corollary 1 *When* $R^2 < \frac{\rho}{\left[\frac{cov(\hat{y}_i^p,\varphi_i)}{var(\hat{y}_i^p)}-\frac{cov(y_i^p,\varphi_i)}{var(y_i^p)}\right]+\rho}$, $\hat{\rho}^{TSTSLS-SMI}$ *is lower than* $\hat{\rho}$.

Proof of Corollary 1 Follows directly from Eq. 8 from the online appendix. □

Corollary 1 states that when the R^2 is lower than some bound, the TSTSLS SMI IGE is a lower bound of the true IGE.

5.2 Cell multiple imputation

Here we investigate another multiple imputation procedure, which we denominate as cell multiple imputation, first used by Jácome et al. (2023). Assume that $\hat{y}_i^p$ can be chosen on a cell built using the predictors in the first stage. For example, suppose that we have education as a predictor for parental earnings. This variable has 18 levels and each of those levels represents a cell. Thus, for each education level, we can obtain a prediction for y_i^p randomly chosen from all the parental earnings associated with that level of education. If we add an additional variable with two levels, we can build 18×2 cells. In the case of continuous variables, these can be coarsened using some criterion. Thus, formally, call $\kappa \in \{1, ..., C\}$ a cell. We can write $\hat{\hat{y}}_i^p$ (the multiple imputation prediction for parent i) as $\hat{\hat{y}}_{i,m}^p(\kappa)$ randomly drawn from the parental earnings belong to a cell κ. We can write this as:

$$\hat{\hat{y}}_{i,m}^p(\kappa) = \hat{y}_{i,m}^p(\kappa) + e_{i,m}, \tag{31}$$

where, $\hat{y}^p_{i,m}(\kappa)$ is the average parental earning in cell κ computed through a TSTSLS procedure, and $e_{i,m} = \hat{\hat{y}}^p_{i,m}(\kappa) - \hat{y}^p_{i,m}(\kappa)$. We define the IGE associated to each sample m as $\tilde{\hat{\rho}}^{TSTSLS}_m$. Then, we can generate M samples with this type of imputation, and we can define the TSTSLS cell multiple imputation (CMI) as:

$$\text{plim}\ \hat{\rho}^{TSTSLS-CMI} = \frac{1}{M}\sum_{m=1}^{M}\frac{cov\left(\hat{\hat{y}}^p_{i,m}(\kappa), y^c_i\right)}{var\left(\hat{\hat{y}}^p_{i,m}\right)}, \tag{32}$$

$$= \frac{1}{M}\sum_{m=1}^{M}\frac{cov\left(\hat{y}^p_i(\kappa), y^c_i\right) + cov\left(e_{i,m}, y^c_i\right)}{var\left(\hat{\hat{y}}^p_{i,m}\right)}, \tag{33}$$

$$= \frac{1}{M}\left(\sum_{m=1}^{M}\hat{\rho}^{TSTSLS}_m \cdot R^2 + \frac{cov\left(e_{i,m}, y^c_i\right)}{var\left(\hat{\hat{y}}^p_{i,m}\right)}\right). \tag{34}$$

Where, $\hat{\rho}^{TSTSLS}_m$ is the TSTSLS estimator for the m imputation. Also, $cov\left(e_{i,m}, y^c_i\right)$ in Eq. 34 can be decomposed into $E(e_{i,m}\cdot y^c_i) - E(e_{i,m})\cdot E(y^c_i) = E(e_{i,m}\cdot y^c_i)$. In this context, we can apply the law of expected iteration, which yields

$$E(e_{i,m}\cdot y^c_i) = E\left[E\left(e_{i,m}\cdot y^c_i|\kappa\right)\right], \tag{35}$$

where $E\left(e_{i,m}\cdot y^c_i|\kappa\right)$ can be different from 0 given that y^c_i also depends on the cell values. Define $\lambda \equiv \frac{\sum_{i=1}^{M}\frac{E(e_{i,m}\cdot y^c_i)}{var\left(\hat{\hat{y}}^p_{i,m}\right)}}{M}$. Given this, we can establish conditions when cell multiple imputation is better than TSTSLS to estimate the IGE.

Corollary 2 *Cell multiple imputation improves the estimation of the IGE compared to baseline TSTSLS if the R^2 of the first stage holds:*

$$\frac{-\left[\frac{cov(\hat{y}^p_i,\varphi_i)}{var(\hat{y}^p_i)} - \frac{cov(y^p_i,\varphi_i)}{var(y^p_i)}\right] + \rho}{\left[\frac{cov(\hat{y}^p_i,\varphi_i)}{var(\hat{y}^p_i)} - \frac{cov(y^p_i,\varphi_i)}{var(y^p_i)}\right] + \rho} - \lambda < R^2. \tag{36}$$

Proof of Corollary 1 Follows from proposition 3 and Eq. 34

Thus, as long as $\lambda \neq 0$ the TSTSLS cell multiple imputation IGE differs from the TSTSLS stochastic multiple imputation IGE. In the next section we test empirically the performance of both multiple imputation methods.

6 Empirical analysis

In this section, we perform an empirical Monte Carlo (EMC) exercise to illustrate our explained-variance-endogeneity equality condition by studying the set of instruments that might reduce TSTSLS biases. We also compare the traditional TSTSLS approach to stochastic multiple imputation and cell multiple imputation. Finally, we examine how the differences between the TSTSLS IGE and OLS IGE change as the number of years used to compute parental earnings increases.

6.1 Administrative data for Chile

We use the database of the Chilean unemployment insurance program (UIP) to obtain sons' and their fathers' labor earnings. That is, we do not include mothers or daughters. Moreover, if a father has two sons, then that father appears twice in our database.[11] All people over the age of 18 who have a fixed-term or ongoing contract in the formal private sector are eligible for the UIP. Enrollment in the program is voluntary for contracts initiated prior to September 2002 and compulsory for contracts established after that date.

Father-son linkages were established through administrative records provided by the Civil Registry Office (CRO). Birth certificates issued by this agency contain information on both the son and the father at the time of birth, thus allowing us to identify and build the pairs of sons and fathers included in the UIP database. In our baseline analysis, the sample of sons is composed of individuals that were 29-34 in 2018, while fathers were 31-66 years old in 2007.

We measure fathers' earnings from 2003 to 2007 and sons' earnings from 2014 to 2018 by computing the five-year average of monthly private-sector earnings. In our baseline sample, we only consider fathers and sons that worked at least six months in the formal private sector and we exclude observations with zero earnings as these individuals could be working either as public employees, in the informal sector, or the formal private sector but not covered by the UIP. Moreover, we only include sons and fathers who earn more than half the minimum wage on average to reduce the potential noise from low earnings observations. We have additional information on fathers' background characteristics. First, we have data on fathers' education, which is divided into 18 categories. Second, we have data on the proportion of months that fathers worked under a permanent contract. Third, our dataset considers 10 types of industries, and it contains the share of months that fathers worked in each industry type.[12]

6.2 Empirical Monte Carlo exercise

In this section we mimick a TSTSLS setting through an Empirical Monte Carlo exercise in which we estimate the IGE through TSTSLS under different sets of predictors. Concretely, we adopt the following steps to evaluate the bias of the TSTSLS, the stochastic MI TSTSLS and the cell MI TSTSLS estimators:

i) We obtain $\hat{\rho} = 0.287$ as the OLS IGE estimate using the linked father-son earnings data.

ii) We take a random subsample of 50,000 (out of 282,122 total links) father-son links information from the baseline sample. We then randomly split this subsample into two subsubsamples, of 25,000 observations each. The first subsubsample plays the role of the auxiliary sample of pseudo-fathers and the second plays the role of the main sample. Then, with the auxiliary sample of pseudo-fathers, we estimate the Mincer equation for the pseudo-fathers:

$$y^{pp} = \delta' z^{pp} + v, \tag{37}$$

where in z^{pp} initially includes fathers' education. We estimate δ' by OLS. We then use the fathers' information in the main sample to impute $\hat{y}^p = \hat{\delta}' z^p$. Additionally, we impute 10 samples of fathers' earnings $\tilde{\hat{y}}^p_{i,m}$ imputed through stochastic imputation and 10 sample of fathers' earnings $\hat{\hat{y}}^p_{i,m}$ imputed through cell imputation.

[11] We focus on sons to be in line with the previous literature (Richey and Rosburg 2018; Bloise et al. 2021).

[12] This is the same sample used in Cortes Orihuela et al. (2023).

iii) We compute $\hat{\rho}^{\text{TSTSLS}}$ by regressing y^c on $\hat{y}^p$ from the main sample. We then compute $\hat{\rho}^{\text{TSTSLS-SMI}}$ by averaging the IGEs estimated through stochastic multiple imputation and compute $\hat{\rho}^{\text{TSTSLS-CMI}}$ by averaging the IGEs estimated through cell multiple imputation.
iv) We repeat ii)-iii) 1,000 times.
v) We repeat i)-iv) five times, but each time changing the set of instruments used to predict fathers' earnings. Concretely, in the second iteration we include education, age, type of contract, and industry type as predictors. In the third loop we remove education. These instruments correspond to variables which might be available in survey data. Then, in order to illustrate and measure the importance of our explained-variance-endogeneity equality condition, we need a variable that is highly predictive but not endogenous enough in the children equation. For this, we perform three more loops in which we include a parental fixed effect (FE) not available in traditional household surveys but that can be easily constructed with our data. To estimate this fixed effect we exploit the panel structure of our data through the following equation:

$$y_{it}^{pp} = \alpha_i + \pi z_{it} + \iota_{it}, \tag{38}$$

where y_{it}^{pp} is the logarithm of the monthly average earnings of father i in year t, α_i is a father fixed effect, z_{it} is a vector of time variant observables, and ι_{it} is a white noise term. After adjusting this model by OLS, we recover the estimated fixed effect associated to each father: $\hat{\alpha}_i$. This new variable can be used as an instrument in a TSTSLS setting.
The advantage of including $\hat{\alpha}_i$ is that this variable is highly predictive for parental earnings. Moreover, due to the Frisch-Waugh-Lovell theorem, we are able to remove a large degree of endogeneity of this variable in the children equation by including parental education, occupation and type of contract in the estimation of $\hat{\alpha}_i$. Thus, $\hat{\alpha}_i$ allows us to illustrate Proposition 2 by exemplifying how the instrument selection condition might work.
Given this, in the fourth iteration we only include $\hat{\alpha}_i$. In the fifth iteration we include $\hat{\alpha}_i$, fathers' education, age, type of contract and type of industry. Lastly, in the final iteration, we remove education.
To separate sets of instruments involving usually available variables to those that rely on the individual fixed effect, we call simulations involving the first three sets of instruments our "first exercise" and iterations using the latter three sets our "second exercise".

6.3 Instrument selection for the traditional TSTSLS

We first comment on the performance of the traditional TSTSLS estimator under alternative sets of instruments. Fig. 3 depicts a summary of the first exercise of our simulation procedure. Concretely, the y-axis plots the ratio of the TSTSLS IGE estimator to the OLS IGE, while the x-axis represents alternative sets of instruments used to impute fathers' earnings. Similarly, Table 1 shows the R^2 of the first stage traditional TSTSLS estimator under these alternative models.

As suggested by Fig. 3, the TSTSLS IGE estimator overstates the OLS IGE, suggesting that the instruments used to impute fathers' earnings are more endogenous than predictive. In fact, in an attempt to reduce the endogeneity of imputed fathers' earnings, we remove education in our last specification. However, as suggested by the third row of Table 1, the removal of this variable comes at a great loss of predictive power, as the R^2 decreases by 59%, and so the overall TSTSLS IGE bias does not change by much.

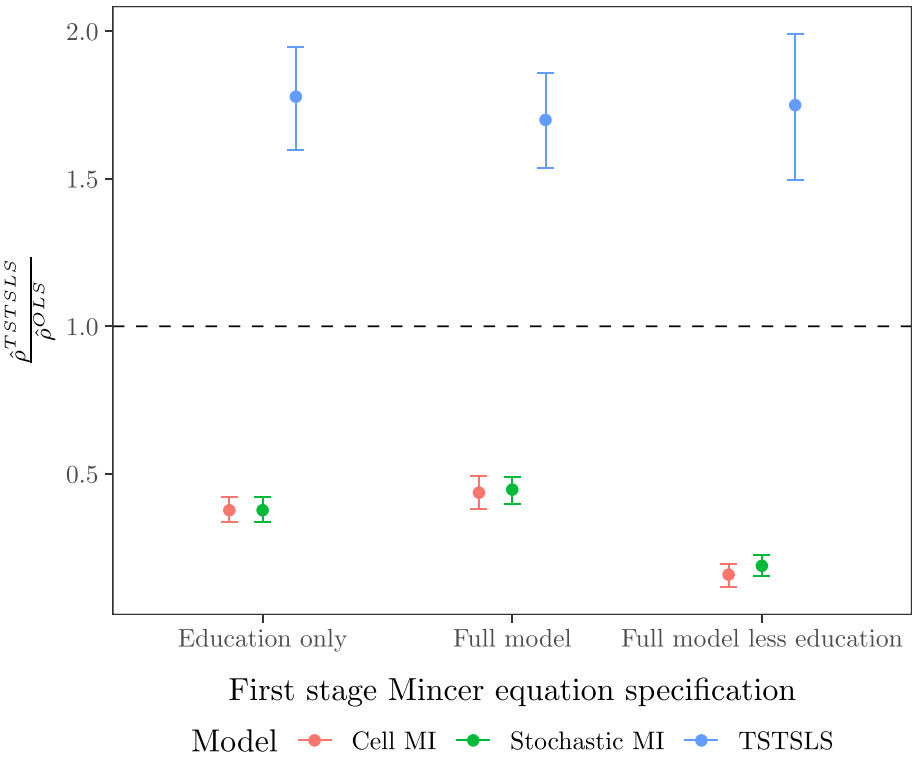

Fig. 3 Performance of the TSTSLS and multiple imputation estimators under alternative sets of instruments. First exercise using usually available variables[†]. [†]*Note*: The y-axis measures the ratio of TSTSLS IGE estimates to OLS IGE estimates, while the x-axis indicates the variables used to estimate fathers' earnings in the first stage. At first we only include fathers' education; we then add fathers' age, type of contract, industry, and education; finally, we remove education. Error bars represent minimum and maximum ratios from the Empirical Monte Carlo exercise, with dots representing average ratio values for 1,000 simulations

Figure 4 depicts the results for our second exercise which involves the use of the individual fixed effect. It can be noted that using this individual fixed effect as the only predictor leads to TSTSLS IGE estimates that understate the OLS IGE, suggesting that this instrument is more

Table 1 R^2, Mincer equation, Alternative set of variables[†]

Instruments	(1) Min.	(2) p25	(4) Mean	(5) p75	(6) Max.
First exercise					
Education	0.189	0.200	0.211	0.223	0.234
Education, Age, Type of contract, Industry	0.242	0.251	0.263	0.274	0.286
Age, Type of contract, Industry	0.094	0.100	0.108	0.116	0.122
Second exercise					
FE	0.482	0.489	0.499	0.509	0.513
FE, Education, Age, Type of contract, Industry	0.800	0.804	0.810	0.816	0.821
FE, Age, Type of contract, Industry	0.623	0.629	0.638	0.646	0.651

Note[†]: This table shows a descriptive statistics on R^2 under alternative set of variables

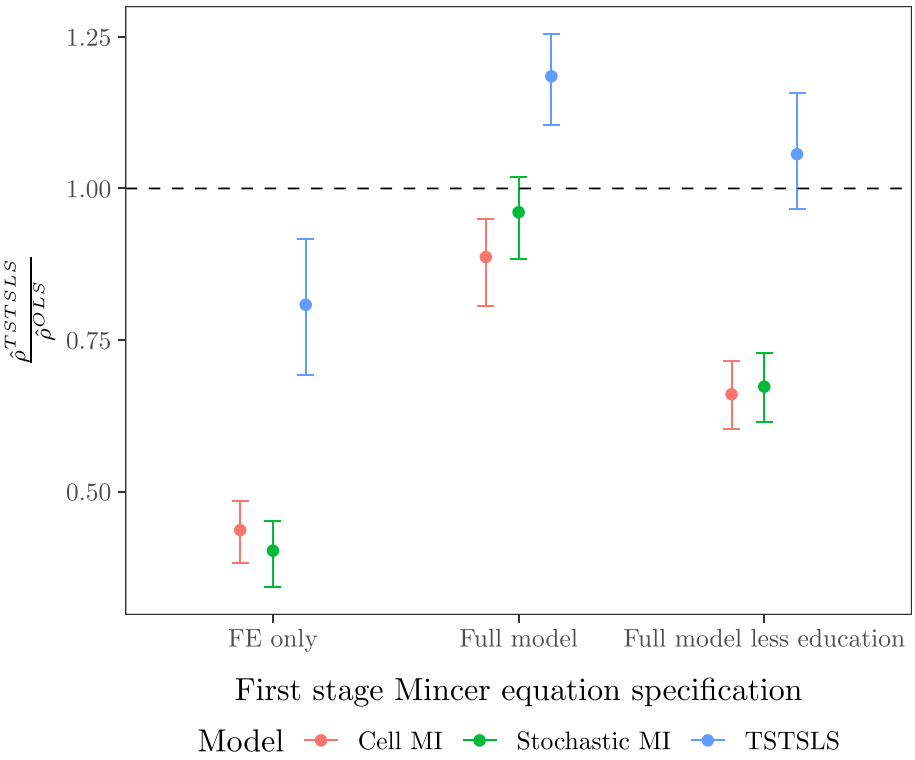

Fig. 4 Performance of the TSTSLS and multiple imputation estimators under alternative sets of instruments. Second exercise using an individual fixed effect†. †*Note*: The y-axis measures the ratio of TSTSLS IGE estimates to OLS IGE estimates, while the x-axis indicates the variables used to estimate fathers' earnings in the first stage. At first we only include an individual fixed effect; we then add fathers' age, type of contract, industry, and education; finally, we remove education. Error bars represent minimum and maximum ratios from the Empirical Monte Carlo exercise, with dots representing average ratio values for 1,000 simulations

predictive than endogenous. Indeed, the R^2 from using this instrument alone is around 0.499, which is greater than for all previous specifications. Once we add back all usually observed variables, we obtain TSTSLS IGE estimates that overstate the OLS IGE, confirming that the added endogeneity of these variables is greater than their predictive power. Once again, in an attempt to reduce the endogeneity of our imputed fathers' earnings, we remove education as a predictor in the last specification. This results in a remarkable decrease in bias which yields TSTSLS estimates that are comparable to the OLS IGE estimate. Albeit the absolute drop in R^2 due to the removal of education is comparable to the drop in the first exercise, the overall higher R^2 of the second exercise implies that the percentage drop in R^2 is lower in the second exercise than in the first. Concretely, in this case the R^2 is reduced by 21%, and so it becomes worthwhile to remove this variable in order to decrease the endogeneity of imputed fathers' earnings. These results suggest that researchers should be strategic when choosing first-stage instruments, as the removal of a particular instrument might yield different results depending on the set of instruments already used. In what follows, we comment on the performance of the multiple imputation estimators.

6.4 Multiple imputation

Figures 3 and 4 also plot the results for the stochastic and cell multiple imputation. First, it can be seen that both procedures produce comparable results, as the min-max error bars for each method intersect for all sets of instruments. Since the asymptotic difference between both methods is given by λ, this suggests that this leftover term is small. We further note that all multiple imputation estimates are significantly lower than the traditional TSTSLS IGE estimates, a result that stems from the fact that two-sample MI estimators are asymptotically close to the traditional TSTSLS estimator but scaled by the first-stage R^2. Moreover, both Figs. 3 and 4 suggest that multiple imputation estimates are below the OLS IGE, with the exception of the scenario which uses all variables including the individual fixed effect (see Fig. 4). This result is consistent with the previous literature, which established that the TSTSLS SMI estimator provides a lower bound for the OLS IGE under plausible assumptions (Cortes Orihuela et al. 2023). Thus, we extend previous results by empirically finding that the cell multiple imputation method used by Jácome et al. (2023) also provides a lower bound for the OLS IGE. In addition to our analysis, we perform a sensibility test regarding measurement error. Those results can be found in the online appendix.

7 Concluding remarks

This paper contends that the TSTSLS IGE differs from the OLS IGE due to two issues: the replication of (i) the variance of unobserved parental earnings and (ii) the endogeneity of unobserved parental earnings in the equation of children's earnings. In this regard, we propose the explained-variance-endogeneity equality condition, which implies that TSTSLS consistency can be achieved even with an imperfect prediction of parental earnings. Moreover, we find that using exogenous instruments to impute parental earnings might not lead to TSTSLS consistency. This paper also analyzes the performance of two-sample multiple imputation techniques to estimate the IGE when parental earnings are missing. Albeit these techniques solve the differences in variance between imputed and actual parental earnings, they do not deal with the replication of the endogeneity of parental earnings in the children's earnings equation. In this sense, this procedure does not consider one of the sources of bias plaguing the TSTSLS.

While we believe that this paper advances the literature on the TSTSLS estimator, it does not examine all relevant issues on the matter. In particular, we identify four avenues for future research. First, scholars should examine extensively what set of instruments yield the best results at replicating the OLS IGE through TSTSLS. As evidenced in this paper, removing parental education as an instrument may be beneficial in some contexts, and so the question of what predictor variables should be used is not a trivial one. Second, recently, there has been interest in the historical study of intergenerational mobility using two-sample procedures (Jácome et al. 2023). However, there has been no formal study on how the estimation of trends in intergenerational mobility might be affected by imputation procedures. Third, the analysis of TSTSLS biases should be expanded to the multivariate case when there are multiple income sources some of which are not observed. This is particularly relevant for developing countries where informal earnings represent a large part of individual's earnings. In such cases, imputation methods are needed to impute earnings but the biases arising from such procedures have not been examined. Fourth, in the paper we have assumed that the covariance between imputed and actual parental log-earnings is asymptotically equivalent to

the variance of imputed parental log-earnings, and so the R^2 of the first-stage is equivalent to the R^2 of imputed parents in the main sample. This implies that the normal equations hold in the main sample. However, this might not be the case in actual two-sample applications. For example, the auxiliary sample of pseudo-parents might not be entirely representative of actual parents, potentially affecting TSTSLS consistency. While there has been research on how heterogeneous samples might impact the estimation of causal effects (Zhao et al. 2019), it remains to be seen how these biases affect the estimation of the IGE. Similarly, overfitting might yield a high R^2 in the first-stage but a low out-of-sample R^2. Albeit prominent research has been done on the issue of overfitting in a TSTSLS setting (Bloise et al. 2021), the exact biases that arise from the violation of the normal equations are yet to be fully understood.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10888-024-09643-8.

Acknowledgements We would like to thank to Claudio Ferraz, Nicole Fortin, David Green, Thomas Lemieux, Kevin Milligan, Lars Osberg, Shane Singh and the participants of the UBC applied economics seminar. We thank the Budget Office and Ministry of Education of Chile for providing access to the data used in this work. Finally, we thank Fabián Abarza for his invaluable research assistance in this project.

Author Contributions All authors work equally in all parts of the paper

Funding No funding was received for this study.

Data Availability The data that support the findings of this study are available from the Chilean government but restrictions apply to the availability of these data, which were used under a confidentiality contract for the current study, and so are not publicly available. Codes are however available from the authors upon request.

Declarations

Competing interests The authors declare no competing interests.

Ethical approval Not applicable.

References

Acciari, P., Polo, A., Violante, G.L.: And yet it moves: intergenerational mobility in Italy. Am. Econ. J. Appl. Econ. **14**(3), 118–63 (2022)

Angrist, J.D., Krueger, A.B.: The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. J. Am. Stat. Assoc. **87**(418), 328–336 (1992)

Björklund, A., Jäntti, M.: Intergenerational income mobility in Sweden compared to the United States. Am. Econ. Rev. **87**(5), 1009–1018 (1997)

Bloise, F., Brunori, P., Piraino, P.: Estimating intergenerational income mobility on sub-optimal data: a machine learning approach. J. Econ. Inequality **19**(4), 643–665 (2021)

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

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

Corak, M., Heisz, A.: The intergenerational earnings and income mobility of Canadian men: evidence from longitudinal income tax data. J. Human Resources, 504–533 (1999)

Cortes Orihuela, J., Díaz, J.D., Gutiérrez Cubillos, P., Troncoso, P.A.: Everything's not lost: revisiting TSTSLS estimates of intergenerational mobility in developing countries. International Tax and Public Finance, 1–29 (2023)

Deutscher, N., Mazumder, B.: Intergenerational mobility across Australia and the stability of regional estimates. Labour Econ. **66**, 101861 (2020)

Inoue, A., Solon, G.: Two-sample instrumental variables estimators. Rev. Econ. Stat. **92**(3), 557–561 (2010)
Jácome, E., Kuziemko, I., Naidu, S.: Mobility for all: Representative intergenerational mobility estimates over the 20th century, Technical report, Working paper (2023). https://elisajacome.github.io/Jacome/historical_mobility.pdf
Jerrim, J., Choi, Á., Simancas, R.: Two-sample two-stage least squares (tstsls) estimates of earnings mobility: how consistent are they?, In: Survey Research Methods, Vol. 10, pp. 85–101 (2016)
Kenedi, G., Sirugue, L.: Intergenerational income mobility in france: a comparative and geographic analysis. Journal of Public Economics **226**, 104974 (2023). https://www.sciencedirect.com/science/article/pii/S0047272723001561
Klevmarken, A.: Missing variables and two-stage least-squares estimation from more than one data set, Technical report, IUI Working Paper (1982)
Nicoletti, C., Ermisch, J.F.: Intergenerational earnings mobility: changes across cohorts in Britain. The BE J. Econ. Anal. Policy **7**(2) (2008)
Richey, J., Rosburg, A.: Decomposing economic mobility transition matrices. J. Appl. Economet. **33**(1), 91–108 (2018)
Rubin, D.B.: Multiple imputation for nonresponse in surveys, Vol. 81, John Wiley & Sons (2004)
Solon, G.: Intergenerational income mobility in the United States. American Econ. Rev., 393–408 (1992)
Zhao, Q., Wang, J., Spiller, W., Bowden, J., Small, D.S.: Two-sample instrumental variable analyses using heterogeneous samples. Stat. Sci. **34**(2), 317–333 (2019)
Zimmerman, D.J.: Regression toward mediocrity in economic stature. American Econ. Rev., 409–429 (1992)