View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Springer - Publisher Connector

Empir Econ (2016) 50:299–316
DOI 10.1007/s00181-015-0941-z

Linear regression with an estimated regressor: applications to aggregate indicators of economic development

Lingsheng Meng · Binzhen Wu · Zhaoguo Zhan

Received: 15 January 2014 / Accepted: 22 January 2015 / Published online: 7 February 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract This study examines the consequences of using an estimated aggregate measure as an explanatory variable in linear regression. We show that neglecting the seemingly small sampling error in the estimated regressor could severely contaminate the estimates. We propose a simple statistical framework to account for the error. In particular, we apply our analysis to two aggregate indicators of economic development, the Gini coefficient and sex ratio. Our findings suggest that the impact of the estimated regressor could be substantially underestimated, when the sampling error is not accounted for.

Keywords Estimated regressor · Sampling error · Gini · Sex ratio · Inequality

JEL Classification C1 · C2 · D6

1 Introduction

Empirical studies often encounter the following situation: A regressor in the linear regression needs to be estimated before it is included in the regression analysis. Such a regressor could be an aggregate measure, which is unavailable but can be estimated with micro data. Examples of an estimated regressor include the widely used Gini coefficient for economic inequality and sex ratio for gender imbalance, see, e.g., Atkinson

L. Meng · B. Wu · Z. Zhan (✉)
School of Economics and Management, Tsinghua University, Beijing 100084, China
e-mail: zhanzhg@sem.tsinghua.edu.cn

L. Meng
e-mail: menglsh@sem.tsinghua.edu.cn

B. Wu
e-mail: wubzh@sem.tsinghua.edu.cn

and Brandolini (2001), Alesina and Angeletos (2005), Edlund et al. (2009), Jin et al. (2011) and Wei and Zhang (2011).

When an estimated regressor is subject to sampling error, the ordinary least squares (OLS) estimator is potentially biased. Nevertheless, the data used to estimate this regressor can be employed to infer the error. With the inferred information, we propose an adjusted version of the OLS estimator, which accounts for sampling error in the estimated regressor. We find that the OLS estimator without accounting for sampling error could severely underestimate the effect of the estimated regressor.

The situation under consideration is closely related to the setup of measurement error or generated regressors (e.g., predicted values or residuals of linear regression as regressors), both of which have been well studied in the existing literature (see, e.g., Hausman 2001; Murphy and Topel 1985). However, the situation considered in this paper and that in existing studies exhibit subtle differences. First, the sampling error associated with an estimated regressor is typically heteroscedastic with a nonzero mean, whereas the classical measurement error is assumed to be homoscedastic with a zero mean. Second, each observation of an estimated regressor is usually computed independent of the other observations, and the method to estimate the regressor might also vary across observations. By contrast, generated regressors typically result from a common functional form that holds across observations. These subtle differences imply that existing methods, such as the classical errors-in-variables estimator and the adjustment in Murphy and Topel (1985), are no longer suitable to correct the sampling error in an estimated regressor.

The sampling error associated with an estimated regressor can be dealt with by the instrumental variable (IV) approach. However, the sampling error of an estimated regressor is often neglected in practice for two reasons. First, finding variables that can serve as valid instruments may be difficult. Second, one may think that the sampling error is small and thus negligible, particularly when the sample size is large. Assuming that neglecting the small sampling error will not severely bias the estimates is sometimes tempting.

In this paper, we show that the cost of neglecting the sampling error in an estimated regressor could be substantial, even when the error is small. The underlying reason is that if the variation of the estimated regressor itself is also small, the seemingly small sampling error could lead to a large difference in the estimates. We illustrate this difference by comparing the OLS estimator with its adjusted version that accounts for the sampling error. The proposed adjustment relies on the data used to estimate the regressor and does not turn to IV or the generalized method of moments (GMM).

We use the Gini coefficient and sex ratio as examples of estimated regressors. A large body of literature in development economics uses Gini as an indicator of economic inequality and sex ratio as a measure of gender imbalance. Although the Gini coefficient and sex ratio are typically estimated by large survey data sets with a small sampling error, their own variation is also small, see, e.g., Deininger and Squire (1996) and Barro (2008) on the relatively small variation of Gini, so their seemingly small sampling error is generally non-negligible. We illustrate the non-negligibility of the sampling error with two empirical examples. Using the same data as in Jin et al. (2011) and Wei and Zhang (2011), we find evidence that the OLS estimator is substantially different from its adjusted version that takes sampling error into account.

For example, using the data in Jin et al. (2011), we find that the OLS point estimate of the effect of Gini increases (in absolute value) by over 170 % when sampling error is accounted for.

Although the Gini coefficient and sex ratio are used as our main examples, the message conveyed here also applies to other aggregate indicators of economic development that are associated with sampling error and that serve as regressors. These aggregate indicators include per capita income, infant mortality, and literacy rate, to name a few. If sampling error of an estimated regressor appears comparable with the variation of the estimated regressor, then the empirical findings related to this regressor generally need careful reexamination, because the estimates can be severely underestimated as a result of ignorance of the sampling error.

The rest of this paper is organized as follows. In Sect. 2, we describe a linear regression model with an estimated regressor, and propose an adjusted version of the OLS estimator that accounts for the sampling error associated with the estimated regressor. Section 3 includes two empirical applications to illustrate the improvement made by the proposed adjustment for the sampling error. Section 4 concludes. Further details and the Monte Carlo evidence are presented in the Appendix.

2 Model and adjustment

2.1 Linear regression with an estimated regressor

Consider a linear regression that captures the relationship of economic variables for N groups:

$$y_i = \alpha_i \cdot \beta + \boldsymbol{\Delta}_i' \boldsymbol{\gamma} + \epsilon_i, \ i = 1, 2, \ldots, N. \tag{1}$$

where y_i is an economic variable of interest, α_i denotes some population measure (such as the Gini coefficient and sex ratio) for the ith group, $\boldsymbol{\Delta}_i$ is a vector of control variables, and ϵ_i is the exogenous error. β and $\boldsymbol{\gamma}$ denote the parameters. In particular, β is the parameter of interest. N is the total number of groups for this model.

2.1.1 Example I: Gini

Such a linear regression model often appears in the vast literature on economic growth and income inequality. In this literature, y_i denotes the economic growth of the ith nation or province/state (or in a panel data setup, the ith intersection of nation and time), whereas income inequality is usually measured by the Gini coefficient, corresponding to α_i in the model above. The parameter β is of interest, see, e.g., Barro (2000).

Besides the effect of income inequality on economic growth, this model is similarly applied to analyze the effect of inequality on consumption, investment, migration, and health, see, e.g., Atkinson and Brandolini (2001), Alesina and Angeletos (2005) and Jin et al. (2011). In all of these studies, the population Gini coefficient is unknown, so empirical researchers have to work with the sample Gini coefficient denoted as $\hat{\alpha}_i$, which is an estimator for the population Gini α_i in the ith group.

2.1.2 Example II: Sex ratio

Such a linear regression is also widely applied in the literature on gender inequality. For example, Wei and Zhang (2011) relate the savings rate in a region to its sex ratio (men over women). y_i denotes the savings rate in region i, and α_i is the sex ratio in this region. Wei and Zhang (2011) hypothesize that β is positive, i.e., high sex ratios lead to high savings rates. Similarly, Edlund et al. (2009) relate high sex ratios to crime rates.

In both Edlund et al. (2009) and Wei and Zhang (2011), the sex ratio in a region is estimated by the genders of individuals sampled in this region, i.e., the estimated sex ratio $\hat{\alpha}_i$ is used as a regressor, instead of the population sex ratio α_i in the empirical analysis.

2.2 Sampling error with a nonzero mean

For the model described by (1), the data for y_i and $\boldsymbol{\Delta}_i$ are generally readily available, but α_i is unknown and needs to be estimated by its sample counterpart $\widehat{\alpha}_i$. If α_i is the population Gini coefficient for nation i, then it is unknown but can be estimated by, e.g., some sampled individual income data from this nation. Similarly, if α_i denotes the sex ratio in region i, it also needs to be estimated by the sampled individuals.

Because $\widehat{\alpha}_i$ differs from α_i as a result of sampling error, we write

$$\widehat{\alpha}_i = \alpha_i + u_i \tag{2}$$

where the difference between $\widehat{\alpha}_i$ and α_i, denoted by u_i, is the sampling error. Consequently, the actual model faced by empirical researchers is as follows:

$$y_i = \widehat{\alpha}_i \cdot \beta + \boldsymbol{\Delta}_i' \boldsymbol{\gamma} + \tilde{\epsilon}_i, \quad \text{where } \tilde{\epsilon}_i = \epsilon_i - u_i \cdot \beta. \tag{3}$$

When $\beta \neq 0$, $\tilde{\epsilon}_i$ is correlated with $\widehat{\alpha}_i$. The estimated regressor $\widehat{\alpha}_i$ thus suffers from endogeneity because of its sampling error, which jeopardizes the OLS estimator for β.

However, neglecting the sampling error u_i associated with the estimated regressor $\widehat{\alpha}_i$ is common practice, particularly when the size of the data used to compute $\widehat{\alpha}_i$ is large. For instance, the standard error associated with the estimated Gini coefficient is seldom reported in empirical studies, and the endogeneity of Gini in the linear regression is often not addressed, e.g., in Deininger and Squire (1998), Kremer and Chen (2002), Alesina and Angeletos (2005) and Jin et al. (2011). The sampling error of sex ratio is similarly neglected in Edlund et al. (2009). Asymptotically, neglecting the sampling error u_i is not completely unreasonable. As the sample used for computing $\widehat{\alpha}_i$ increases, u_i will decrease to zero, so the OLS estimator for β is expected to be consistent under regularity conditions. However, we will highlight in this paper that the cost of neglecting the sampling error u_i in (3) can be high in finite sample applications, even when u_i appears small. We then propose a method to adjust for this error.

Note that our model of (1)–(3) effectively describes an errors-in-variables problem, or a measurement error problem. That is, the unknown regressor α_i is contaminated by measurement error u_i, according to (2). Consequently, the empirical findings are contaminated if the measurement error is not treated, see, e.g., Hausman (2001).

Different from the classical errors-in-variables model, where the measurement error is assumed to be homoscedastic with a zero mean, our model allows the sampling error u_i to be heteroscedastic with a nonzero mean. This characteristic corresponds to two facts. First, the estimator $\widehat{\alpha}_i$ for α_i could be biased, and this induces the nonzero mean of u_i. Second, for a different group, its population measure α_i may be estimated by $\widehat{\alpha}_i$ with a different sample size, which naturally induces possibly different variances of the sampling error u_i, for $i = 1, 2, \ldots, N$.

Therefore, our model can be viewed as a natural extension of the classical measurement error model. Assuming that the estimator $\widehat{\alpha}_i$ has finite variance σ_i^2, we can thus rewrite its associated sampling error u_i as follows:

$$u_i = b_i + \sigma_i \tau_i \tag{4}$$

where $b_i \equiv \mathbb{E}(\widehat{\alpha}_i) - \alpha_i$ denotes the bias of the estimator $\widehat{\alpha}_i$, and τ_i is a random variable with a zero mean and unit variance. In the classical setup where u_i is homoscedastic with a zero mean, (4) reduces to $u_i = \sigma_u \tau_i$, where σ_u^2 is the same variance of u_i, for $i = 1, 2, \ldots, N$.

Furthermore, unlike the classical measurement error, whose distribution is typically unknown, the distribution of sampling error can be derived or approximated. The reason is that deriving or approximating the distribution of the estimator $\widehat{\alpha}_i$ is usually possible. Consequently, we can infer the distribution of u_i because $u_i = \widehat{\alpha}_i - \alpha_i$, and α_i is a fixed parameter for a given i. When the data used for computing $\widehat{\alpha}_i$ is available, we may also use the data to approximate its associated bias b_i and standard error σ_i. For instance, if $\widehat{\alpha}_i$ stands for the estimated Gini coefficient, Deltas (2003) and Langel and Tillé (2013) show how to use individual income data to approximately derive the bias and standard error of $\widehat{\alpha}_i$.

2.3 Relation to existing literature

Before proceeding to the adjustment for the sampling error of the estimated regressor, we first show how our model is related to the existing econometric literature.

So far, we have explained that our model is not fully nested by the classical measurement error model, although they are closely related. Similarly, our model is also closely related to, but not nested by, the existing literature of generated regressors, see, e.g., Murphy and Topel (1985) and Hoffman (1987) for an early discussion of this topic.

From the perspective of generated regressors, our model corresponds to a two-step procedure: In the first step, the regressor $\widehat{\alpha}_i$ is generated or estimated; in the second step, the generated or estimated regressor $\widehat{\alpha}_i$ is included in the regression analysis.[1]

[1] Two approaches coexist to handle the problem induced by generated regressors. One approach (see, e.g, Murphy and Topel 1985) is to adjust the variance of the OLS estimator in the second step to account for the

We assume that the first-step estimation in our model is independent of the second-step main regression, i.e., the sampling error u_i is independent of the dependent variable y_i, the controls $\boldsymbol{\Delta}_i$ and the structural error ϵ_i.

However, our model of (1)–(4) does not fully fit into the existing literature of generated regressors. As in Murphy and Topel (1985) and Hoffman (1987), generated regressors typically result from a common functional form with certain (unknown) parameters. By contrast, in our model of estimated regressors, each $\widehat{\alpha}_i$ is computed independently, for $i = 1, 2, \ldots, N$; in addition, we allow the way of computing $\widehat{\alpha}_i$ to vary, i.e., $\widehat{\alpha}_i$ and $\widehat{\alpha}_j$ may result from two different procedures, when $i \neq j$. In other words, a common functional form with the same parameters that can describe how $\widehat{\alpha}_i$ is generated does not exist,[2] for $i = 1, 2, \ldots, N$. Based on the argument above, the existing methods from the literature on generated regressors are not suitable to account for the independent and heterogeneous sampling error in our model.

To handle the endogeneity problem described in (1)–(4), a potential solution is to use an instrumental variable for $\widehat{\alpha}_i$ and conduct the IV estimation. Generally, when more instrumental variables or identification conditions are available, the GMM approach can also be adopted. However, the availability of a good instrumental variable in every empirical application is not guaranteed. Furthermore, even when such an instrumental variable is available, it may suffer from the so-called weak instrument problem examined in Stock et al. (2002), who warn that the IV and the GMM estimator are still unreliable if the statistical quality of instruments is weak. Consequently, having a method other than IV/GMM is useful to bypass the endogeneity because of sampling error.

To summarize, we have described a model for the sampling error associated with an estimated regressor, as in (1)–(4). Although the model setup appears similar to that of the measurement error or generated regressors, some subtle differences exists, so that the current methods in the literature of measurement error and generated regressors cannot be directly applied to our model. In addition, we do not intend to resolve the sampling error problem by turning to IV or GMM, both of which call for extra requirements. In the remaining part of the paper, we show that if the bias and standard error associated with the estimated regressor $\widehat{\alpha}_i$ can be approximated, then the OLS estimator for β can be directly adjusted to account for the sampling error of $\widehat{\alpha}_i$. As detailed below, the proposed adjustment is a modified version of the classical errors-in-variables estimator.

Footnote 1 continued
sampling error associated with the generated regressors in the first-step estimation. The other approach (see, e.g, Hoffman 1987) is to consider a system that includes both the first and the second step and simultaneously apply some estimation method (e.g., generalized least squares, maximum likelihood) to the system of both steps to avoid the error associated with generated regressors.

[2] Although the formula or expression to compute $\widehat{\alpha}_i$ can be the same for every i, the data used for the formula are usually drawn from a different underlying distribution for a different i. For example, even if the way of computing Gini is the same for every country, the income data of different countries are drawn from different income distributions, which cannot be described by the same functional form with the same parameters.

2.4 Adjustment for the sampling error

We start the econometric discussion with the OLS estimator of β, denoted by $\hat{\beta}_{\text{OLS}}$:

$$\hat{\beta}_{\text{OLS}} = \frac{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{Y}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}} \tag{5}$$

where $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_N)'$, $\mathbf{M}_{\boldsymbol{\Delta}} = \mathbf{I} - \boldsymbol{\Delta}(\boldsymbol{\Delta}'\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'$, $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \ldots, \boldsymbol{\Delta}_N)'$, $\mathbf{I}$ is the identity matrix, and $\mathbf{Y} = (y_1, y_2, \ldots, y_N)'$.

To account for the sampling error associated with the estimated regressor $\hat{\alpha}_i$, we propose an adjusted version of $\hat{\beta}_{\text{OLS}}$, denoted by $\hat{\beta}^{\text{adj}}_{\text{OLS}}$ below:

$$\hat{\beta}^{\text{adj}}_{\text{OLS}} = \frac{\hat{\beta}_{\text{OLS}}}{1 - \frac{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \hat{\mathbf{b}} + \hat{\boldsymbol{\sigma}}' \hat{\boldsymbol{\sigma}}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}}}} \tag{6}$$

where $\hat{\mathbf{b}} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_N)'$, $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_N)'$. $\hat{b}_i$ and $\hat{\sigma}_i$ are the approximated bias and standard error of $\widehat{\alpha}_i$, respectively.

For a clear illustration of (6), let us consider a simple case that corresponds to the classical measurement error setup. Assume that the mean of u_i is zero, and its variance $\sigma_i^2 = \sigma_u^2$, for $i = 1, 2, \ldots, N$. This simplification thus requires that the estimated regressor $\widehat{\alpha}_i$ is unbiased with the same variance σ_u^2 for each i. In this simplified case, (6) reduces to the classic errors-in-variables estimator with $\hat{\mathbf{b}} = \mathbf{0}$ and $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_u, \hat{\sigma}_u, \ldots, \hat{\sigma}_u)'$. In other words, (6) is a modified version of the classical errors-in-variables estimator, and the modification corresponds to the heterogenous feature of sampling error associated with the estimated regressor.

Notably, the ratio $\frac{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \hat{\mathbf{b}} + \hat{\boldsymbol{\sigma}}' \hat{\boldsymbol{\sigma}}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}}$ in (6) helps explain why the possibly small sampling error of $\widehat{\alpha}_i$ may not be negligible. Although the bias and standard error of the estimated regressor may be small, the cross-sectional variation of $\widehat{\alpha}_i$ after the control variables are projected out, $\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}$, may also be small. If so, $\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \hat{\mathbf{b}} + \hat{\boldsymbol{\sigma}}' \hat{\boldsymbol{\sigma}}$ and $\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}$ may probably have similar magnitudes, which induce the malfunction of $\hat{\beta}_{\text{OLS}}$.

The difference between (5) and (6) indicates that $\hat{\beta}^{\text{adj}}_{\text{OLS}}$ takes the sampling error into account, whereas $\hat{\beta}_{\text{OLS}}$ does not. Consequently, $\hat{\beta}^{\text{adj}}_{\text{OLS}}$ is expected to outperform $\hat{\beta}_{\text{OLS}}$, particularly when the sampling error is sizeable. When the sampling error is negligible, i.e., $\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \hat{\mathbf{b}} + \hat{\boldsymbol{\sigma}}' \hat{\boldsymbol{\sigma}}$ has a much smaller magnitude than $\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}$, $\hat{\beta}^{\text{adj}}_{\text{OLS}}$ is similar to $\hat{\beta}_{\text{OLS}}$. For statistical inference, we also provide an expression for the standard error of $\hat{\beta}^{\text{adj}}_{\text{OLS}}$:

$$\text{s.e.}\left(\hat{\beta}^{\text{adj}}_{\text{OLS}}\right) = \frac{\left\{\left(\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}\right)^{-1} \left[\left(\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \hat{\boldsymbol{\epsilon}}\right)' \left(\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \hat{\boldsymbol{\epsilon}}\right)\right] \left(\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}\right)^{-1}\right\}^{1/2}}{1 - \frac{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \hat{\mathbf{b}} + \hat{\boldsymbol{\sigma}}' \hat{\boldsymbol{\sigma}}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}}}} \tag{7}$$

which is a scaled version of the standard error of $\hat{\beta}_{\text{OLS}}$, with $\hat{\boldsymbol{\epsilon}} = \mathbf{M}_{\boldsymbol{\Delta}}[\mathbf{Y} - (\widehat{\boldsymbol{\alpha}} - \hat{\mathbf{b}})\hat{\beta}^{\text{adj}}_{\text{OLS}}]$. Similarly, when sampling error is negligible, this standard error reduces to the square root of the classical variance estimator of White (1980).

For brevity, the derivation of $\hat{\beta}_{\text{OLS}}^{\text{adj}}$, as well as the Monte Carlo evidence for its validity and usefulness, is shown in the Appendix.

3 Application

In this section, we use two empirical examples to show that the proposed adjustment in (6) can cause a substantial difference, once the seemingly small sampling error is accounted for. In particular, we choose the Gini coefficient and sex ratio as examples of estimated regressors. This choice is motivated by the sizeable literature on economic inequality and gender imbalance, where Gini and sex ratio are widely used.

3.1 Application I: Gini coefficient

As the leading measure of economic inequality, the Gini coefficient widely serves as a regressor in empirical studies. For example, Barro (2000, 2008) relate a nation's economic growth to its Gini coefficient; Alesina and Angeletos (2005) study whether Gini and social belief affect tax and welfare policies; and Jin et al. (2011) argue that high inequality measured by Gini induces less consumption.

However, the accuracy of Gini coefficient used in empirical studies has long been under doubt. Both instrumental variable estimation and the generalized method of methods have been adopted to address the endogeneity of Gini, see, e.g., Forbes (2000) and De La Croix and Doepke (2003). Nevertheless, dealing with the endogeneity of Gini in linear regression analysis is not commonplace yet, and the measurement error of Gini appears to still be ignored in most empirical studies. For instance, the endogeneity of Gini is not addressed in Deininger and Squire (1998), Kremer and Chen (2002), Alesina and Angeletos (2005) and Jin et al. (2011).

The common ignorance of the measurement error of Gini coefficient could result from the belief that this error is small and thus negligible. However, as suggested in the previous section, even the small sampling error of Gini could severely contaminate empirical findings, particularly when the variation of Gini is also small.[3] Given that the sampling error is among various errors that can contaminate Gini, if the sampling error itself can severely contaminate empirical findings, then it implies that the measurement error of Gini generally deserves serious consideration in future studies.

We use the existing methods in the broad literature to compute the Gini coefficient, as well as its associated bias and standard error. α_i now denotes the population Gini coefficient for income (or wealth, expenditure, etc.) inequality in the ith group (or nation, region, etc.), which is defined as twice the area between the 45°-line and the Lorenz (1905) curve. Mathematically, α_i can be written as (see, e.g., Langel and Tillé 2013):

$$\alpha_i = \frac{2}{\mu_i} \int_0^\infty x F_i(x) \mathrm{d}F_i(x) - 1 \tag{8}$$

[3] e.g., The standard error of Gini in Barro (2008) is reported to be around 0.10, before control variables are projected out.

where $F_i(x)$ is the cumulative distribution function (c.d.f.) of income in the ith group, $\mu_i = \int_0^\infty x \mathrm{d}F_i(x)$.

With some random sample of income drawn within the ith group, the commonly used expression for estimating the population Gini coefficient α_i is as follows (see, e.g., Sen 1973; Ogwang 2000):

$$\widehat{\alpha}_i = \frac{2\sum_{j=1}^{n_i} j x_{ij}}{n_i \sum_{j=1}^{n_i} x_{ij}} - \frac{n_i + 1}{n_i} \tag{9}$$

where $\widehat{\alpha}_i$ denotes the (estimated) sample Gini coefficient of the ith group, based on the n_i observations of income, $x_{i1} \leq x_{i2} \leq \cdots \leq x_{in_i}$, x_{ij} is the jth observation of income from the ith group after sorting. The expression of (9) results from replacing the population c.d.f. in (8) with its sample counterpart.

The bias associated with $\widehat{\alpha}_i$ is known to have the leading term $-\alpha_i/n_i$ (see, e.g., Deltas 2003; Davidson 2009), so it can be approximated by

$$\hat{b}_i = -\frac{\widehat{\alpha}_i}{n_i - 1} \tag{10}$$

The standard error $\hat{\sigma}_i$ associated with $\widehat{\alpha}_i$ is often derived by the jackknife method:

$$\hat{\sigma}_i = \left[\frac{n_i - 1}{n_i} \sum_{j=1}^{n_i} \left(\widehat{\alpha}_{i(j)} - \overline{\widehat{\alpha}}_{i(\cdot)} \right)^2 \right]^{1/2} \tag{11}$$

where $\widehat{\alpha}_{i(j)}$ denotes the sample Gini coefficient computed after the jth observation in the ith group is deleted, and $\overline{\widehat{\alpha}}_{i(\cdot)} = \frac{1}{n_i}\sum_{j=1}^{n_i} \widehat{\alpha}_{i(j)}$, see, e.g., Sandström et al. (1988), Ogwang (2000), Modarres and Gastwirth (2006) and Langel and Tillé (2013) for further discussions on the jackknife method for Gini.

To compute $\hat{b}_i$ and $\hat{\sigma}_i$, we need the data set that contains the individual income. This requirement, however, significantly limits our choices of the empirical example, because recovering all the income data used to compute the Gini coefficient that appears in empirical studies is almost impossible. For instance, if we conduct a cross-country study as in Barro (2000, 2008), then we would need to have individual income data used to compute the Gini coefficient for each country. Although such income data might be available, the reliability and comparability of cross-country data sets are under doubt, as stated by Atkinson and Brandolini (2001).

Considering the above reasons, we choose Jin et al. (2011) to illustrate the proposed adjustment in this paper. Instead of computing the Gini coefficient for each country, Jin et al. (2011) use the income data in China to compute the Gini for each peer group that is defined by the interaction of province and age group. The availability of income data to compute Gini in Jin et al. (2011) thus makes our adjustment of the OLS estimator feasible.

To quickly illustrate why the sampling error of Gini might be non-negligible in Jin et al. (2011), we present the summary statistics of the sample Gini and its associated

Table 1 Summary statistics—Gini

	Mean	SD	Min	Max
$\widehat{\alpha}_i$	0.277	0.041	0.116	0.396
$\hat{b}_i$	−0.001	0.001	−0.011	−0.000
$\hat{\sigma}_i$	0.014	0.006	0.006	0.087

$\widehat{\alpha}_i$ stands for the sample Gini coefficient; $\hat{b}_i$ is the estimated bias associated with $\widehat{\alpha}_i$; and $\hat{\sigma}_i$ is the standard error of sampling error associated with $\widehat{\alpha}_i$. The data are the same as those in Jin et al. (2011) (the benchmark result reported in the second column of their Table 1)

Table 2 Regressing consumption on Gini

	Jin et al. (2011)	(I) $\hat{\beta}_{\text{OLS}}$	(II) $\hat{\beta}^{\text{adj}}_{\text{OLS}}$
β	−0.387	−0.238	−0.660
s.e.	(0.121)	(0.056)	(0.156)

The estimate and standard error for β, the parameter of Gini in Jin et al. (2011) (the second column of their Table 1), are re-calculated in two ways: (I) $\hat{\beta}_{\text{OLS}}$, as in (5), and (II) $\hat{\beta}^{\text{adj}}_{\text{OLS}}$, as in (6). For brevity, we omit the estimation outcome of control variables

bias and standard error in Table 1, with the use of the same data for a benchmark result reported in Jin et al. (2011). Two numbers are of particular interest in Table 1. First, the variation of Gini used in Jin et al. (2011) is not very large, as indicated by the reported standard deviation of 0.041. Second, the sampling error associated with the sample Gini is not very small, as indicated by the reported mean 0.014 for the standard error of the sampling error. These two numbers are thus comparable in magnitude. In addition, once control variables are projected out, the variation of Gini is expected to be further reduced: In fact, the standard deviation of Gini will reduce to 0.019 after controls are projected out, and this value is only slightly above 0.014. Consequently, the OLS estimate using the data of Jin et al. (2011) is likely to be severely distorted.

The benchmark regression result reported in Jin et al. (2011) is replicated and presented in the first column of Table 2. For this regression, the dependent variable is the log consumption of the peer group, and the Gini of the peer group is the regressor of interest, whereas the control variables include age, family size, and income, among other variables, see Jin et al. (2011) for further details. To be consistent with our model setup, we do not consider the potential measurement error problem of the dependent variable or control variables. Furthermore, the estimation results of control variables are not included in Table 2, because our interest lies in β, the coefficient of Gini. The reported result for β in the first column of Table 2 is the same as that in Jin et al. (2011), where the estimate of β is roughly −0.387 with s.e. 0.121, by our replication.

The estimation conducted in Jin et al. (2011), however, uses the weight and cluster option[4] for the linear regression analysis. For our purpose of comparing the OLS

[4] In Jin et al. (2011), peer groups are weighted by size, and clustered by province and age.

estimator with its adjusted version, we simply re-estimate their model with the classical OLS method without any further option. The outcome by OLS is reported in Column (I) of Table 2, and it is qualitatively consistent with the results reported in Jin et al. (2011): The OLS estimate of β is found to be negative and significantly different from zero, so a high degree of inequality seems to suggest low consumption, as stated in Jin et al. (2011). However, neither the estimation in Jin et al. (2011) nor the classical OLS method takes the sampling error of Gini into consideration, so the corresponding empirical findings are under doubt.

Column (II) $\hat{\beta}_{\mathrm{OLS}}^{\mathrm{adj}}$ of Table 2 presents the adjusted OLS outcome, with the use of our proposed method to account for the sampling error of Gini. As expected, we find that the point estimate of β after adjustment is much larger than its OLS counterpart in absolute value. The difference between Column (I) and (II) conveys the main message of this paper that ignoring the sampling error of estimated regressors is not cost free, even if the sampling error appears small. If we compare the point estimate -0.238 in Column (I) with its adjusted counterpart -0.660 in Column (II), then the relative difference is found to exceed 170 %. In other words, in this example, taking the sampling error of Gini into consideration increases the OLS estimate by more than 170 % in absolute value. This change in the OLS estimate is sizeable, especially if the estimate is adopted for economic policymaking. In addition, the point estimate -0.238 in Column (I) does not lie in the 95 % confidence interval of β as implied by Column (II), so the resulting difference from the adjustment for the sampling error of Gini also appears significant.

Note that our sole objective of adopting Jin et al. (2011) as an example is to illustrate the substantial change made from the adjustment of the sampling error of Gini. Other than this objective, we do not intend to make any other point out of this example: e.g., we do not propose that reducing Gini by 0.01 will increase consumption by approximately 0.66 %, as Column (II) of Table 2 seems to suggest. Overall, this example indicates that the seemingly small sampling error is not necessarily negligible, and our proposed adjustment could make a substantial difference.

3.2 Application II: Sex ratio

We now turn to another example, where sex ratio serves as the leading regressor in the regression analysis. α_i now stands for the sex ratio in the ith group, and $\hat{\alpha}_i$ is the computed sex ratio based on observations drawn from the ith group.

Specifically, for the ith group, if the fraction of men is denoted by p_i, then the sex ratio in this group is

$$\alpha_i = \frac{p_i}{1 - p_i} \tag{12}$$

Suppose n_i individuals are sampled from the ith group, with $n_{i,m}$ men and $n_i - n_{i,m}$ women, then the sample sex ratio is

$$\hat{\alpha}_i = \frac{n_{i,m}}{n_i - n_{i,m}} \tag{13}$$

By Taylor's expansion, the bias of $\hat{\alpha}_i$ can be approximated by:

$$\hat{b}_i = \frac{n_{i,m}}{\left(n_i - n_{i,m}\right)^2} \tag{14}$$

Furthermore, by the Delta Method, the standard error of $\hat{\alpha}_i$ can be approximated by:

$$\hat{\sigma}_i = \left[\frac{n_i n_{i,m}}{\left(n_i - n_{i,m}\right)^3}\right]^{1/2} \tag{15}$$

To illustrate how our proposed adjustment can outperform the unadjusted OLS estimator, we consider two estimators for sex ratio in this application for a clear illustration. The first estimator is the same as that used in Wei and Zhang (2011), and it is computed by the full sample of the 2,000 Population Census in China, with around 10^7 observations to calculate the sex ratio in each group. By contrast, the other estimator for sex ratio is computed by the 0.5 % sample of the same census, with around 5×10^4 observations to compute each sex ratio. Consequently, the first estimator is expected to be very close to the population sex ratio, whereas the sampling error problem is expected to be more severe for the second estimator, than for the first one; the error problem also results from reasonably large data sets.

Table 3 presents the summary statistics of the two sample sex ratios and their associated bias and standard error. As expected, Table 3 shows that if the sex ratio results from the 0.5 % sample (Panel B), then the sampling error problem is likely to be severe. For example, in Panel B, the variation of sex ratio is not very large (standard deviation ≈ 0.06), whereas the sampling error associated with the sample sex ratio is not very small (e.g., the reported mean is around 0.01 for its standard deviation). These two numbers are thus comparable. By contrast, Panel A indicates that the sampling error problem is negligible under the full sample.

Table 3 Summary statistics—sex ratio

	Mean	SD	Min	Max
Panel A (*full sample of 2,000 Census*)				
$\widehat{\alpha}_i$	1.075799	0.048227	0.925735	1.227102
$\hat{b}_i$	4.46e−07	5.41e−07	6.34e−08	3.21e−06
$\hat{\sigma}_i$	0.000852	0.000449	0.000350	0.002583
Panel B (0.5 % *of full sample*)				
$\widehat{\alpha}_i$	1.059553	0.058870	0.916331	1.225672
$\hat{b}_i$	0.000097	0.000121	0.000014	0.000713
$\hat{\sigma}_i$	0.012432	0.006671	0.005129	0.038288

$\widehat{\alpha}_i$ stands for the sample sex ratio; $\hat{b}_i$ is the estimated bias associated with $\widehat{\alpha}_i$; and $\hat{\sigma}_i$ is the standard error of sampling error associated with $\widehat{\alpha}_i$. The data are from Wei and Zhang (2011), and they correspond to the results reported in their Table 14

Table 4 Regressing savings rate on sex ratio

We use two ways to recalculate the estimate and standard error for β, the parameter of sex ratio in six specified linear models of Wei and Zhang (2011) (Column 1–6 of their Table 14): (I) $\hat{\beta}_{\text{OLS}}$, where the sex ratio is from the 0.5 % sample of the 2,000 Population Census, instead of the sex ratio from the full sample in Wei and Zhang (2011); (II) $\hat{\beta}^{\text{adj}}_{\text{OLS}}$, the adjusted version of $\hat{\beta}_{\text{OLS}}$. For brevity, we omit the estimation outcome of control variables

Specification	Wei and Zhang (2011)	(I) $\hat{\beta}_{\text{OLS}}$	(II) $\hat{\beta}^{\text{adj}}_{\text{OLS}}$
1.	0.282	0.219	0.272
	(0.052)	(0.041)	(0.046)
2.	0.576	0.261	0.399
	(0.178)	(0.142)	(0.137)
3.	0.735	0.544	0.667
	(0.154)	(0.117)	(0.081)
4.	0.282	0.219	0.270
	(0.051)	(0.041)	(0.046)
5.	0.320	0.232	0.302
	(0.062)	(0.046)	(0.055)
6.	0.239	0.161	0.216
	(0.068)	(0.052)	(0.063)

Table 4 presents the linear regression outcome, which corresponds to six model specifications in Wei and Zhang (2011) (see Column 1–6 in their Table 14 for details). The dependent variable is the savings rate, whereas the sex ratio is the leading regressor. The six specifications differ in the choice of control variables. The first column of Table 4 by our replication is the same as the outcome reported in Wei and Zhang (2011), where sex ratio results from the full sample of the census.

For Column (I) $\hat{\beta}_{\text{OLS}}$ of Table 4, we replace the sex ratio used in Wei and Zhang (2011) with its counterpart based on the 0.5 % sample of the 2,000 Census. Under this alternative sex ratio, all OLS estimates of β are found to decrease by at least 20 %.[5] This decrease should not be surprising, because the sampling error tends to bias the OLS estimator toward zero. However, the exercise in Column (I) suggests that the impact of gender imbalance could be severely underestimated in empirical studies where each sex ratio is estimated by around 5×10^4 or fewer observations, see, e.g., Angrist (2002), Edlund et al. (2009), if the sampling error is ignored.

Column (II) $\hat{\beta}^{\text{adj}}_{\text{OLS}}$ of Table 4 presents the adjusted OLS outcome, with the use of our proposed method to account for the sampling error of the sex ratio used for Column (I). As expected, we find that the point estimate of β after adjustment is much larger than its OLS counterpart, and the relative improvement is roughly 20–50 %. In particular, the adjusted outcome in Column (II) is comparable with the result in the first column of Table 4 by Wei and Zhang (2011), and it does not suffer from a severe sampling error.[6]

Overall, Table 4 shows that our empirical framework that adjusts for the sampling error works as expected. When the sampling error of sex ratio is sizeable, the adjusted

[5] We repeatedly randomly draw the 0.5 % sample from the census data 10 times, and the reported numbers are the resulted sample averages.

[6] Our adjusted estimates in Column (II) appear close to but slightly smaller than those reported in Wei and Zhang (2011), and this difference could be caused by, e.g., the 0.5 % sample we used not being an ideal representative of the full sample used in Wei and Zhang (2011).

OLS estimator tends to converge to the baseline estimate for which the sampling error is negligible. Nevertheless, for both applications, we emphasize that we do not claim that our adjusted estimates are free of bias. Strictly speaking, in both models, the Gini coefficient and sex ratio may suffer from various sources of endogeneity, whereas our proposed adjustment only targets the sampling error. Our proposed strategy works best when the sampling error is the single most important source of bias in the aggregate indicator. Conceivably, when the regressor is contaminated with other important sources of bias, e.g., omitted variables, a formal identification strategy (e.g., instrumental variables) is needed to remove all the biases.[7]

4 Conclusion

This study targets a common practice in empirical studies: estimate an unknown regressor with large survey data sets, then include the estimated regressor in the linear regression analysis without accounting for its sampling error. A seemingly reasonable argument for neglecting the sampling error associated with the estimated regressor is that, this error is small, because the data set used to estimate the regressor is large.

We demonstrate in this study that even when sampling error is small, neglecting it may still severely contaminate the regression analysis if the variation of the estimated regressor is also small. We propose an adjustment to account for this error. The proposed adjustment is a modified version of the classical errors-in-variables estimator, because the sampling error is heteroscedastic with a nonzero mean. We use the Gini coefficient and sex ratio as two examples of estimated regressors, and we show that their sampling error is generally non-negligible, by presenting evidence that the OLS estimator may substantially change after the seemingly small sampling error is accounted for.

To conclude, this study highlights that the sampling error of estimated regressors deserves serious consideration, even when these regressors are estimated by large data sets. In addition, the sampling error can be easily accounted for without extra requirements, as long as the data sets used to estimate regressors are available. Alternatively, if bias and the standard errors associated with the estimated regressors are reported in practice, the sampling error can also be addressed without accessing the full data sets. From an empirical perspective, this study also suggests that the existing findings relating the Gini coefficient or sex ratio to other economic variables should be taken with caution, if the measurement error problem is not treated. The real effect of economic inequality or gender imbalance could be much stronger than that reflected by the OLS estimator, if this estimator is not adjusted for measurement error.

Acknowledgments We thank the editor and two anonymous referees for helpful comments and suggestions. Lingsheng Meng acknowledges the financial support from the National Natural Science Foundation of China (Project No. 71303131).

[7] For example, the instrumental variables estimates for β in Wei and Zhang (2011) are much larger (between 0.61 and 1.17 in their Table 16) than our adjusted OLS estimates, $\hat{\beta}_{\text{OLS}}^{\text{adj}}$, reported in Table 4, which indicates that there are other sources of bias other than sampling error in the sex ratio.

5 Appendix

The derivation of $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ is presented here, together with some Monte Carlo evidence.

5.1 $\hat{\beta}_{\text{OLS}}^{\text{adj}}$

For convenience, assume all finite moments of random variables exist, and observations are i.i.d.. Note that

$$\hat{\beta}_{\text{OLS}} = \frac{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{Y}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}}$$

Plug $\mathbf{Y} = \boldsymbol{\alpha}\beta + \boldsymbol{\Delta}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, the vector/matrix version of Eq. (1) with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)'$, into the equation above, we obtain

$$\hat{\beta}_{\text{OLS}} = \frac{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \boldsymbol{\alpha}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}} \beta + \frac{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \boldsymbol{\epsilon}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}} = \left(1 - \frac{\boldsymbol{\alpha}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{U} + \mathbf{U}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{U}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}}\right) \beta + o_p(1)$$

where $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \mathbf{U}$, $\mathbf{U} = (u_1, u_2, \ldots, u_N)'$, and $\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \boldsymbol{\epsilon} / \widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}$ is of order $o_p(1)$.

The vector notation of (4) reads $\mathbf{U} = \mathbf{b} + \boldsymbol{\tau} \odot \boldsymbol{\sigma}$, where $\mathbf{b} = (b_1, b_2, \ldots, b_N)'$, $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_N)'$, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N)'$, $\odot$ denotes element-by-element multiplication. To approximate $\hat{\beta}_{\text{OLS}}$, we write $\boldsymbol{\alpha}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{U} / N$ and $\mathbf{U}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{U} / N$ as follows, after substituting $\mathbf{b} + \boldsymbol{\tau} \odot \boldsymbol{\sigma}$ for $\mathbf{U}$:

$$\frac{\boldsymbol{\alpha}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{U}}{N} = \frac{\boldsymbol{\alpha}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{b}}{N} + o_p(1)$$

where $\boldsymbol{\alpha}' \mathbf{M}_{\boldsymbol{\Delta}} (\boldsymbol{\tau} \odot \boldsymbol{\sigma}) / N$ is of order $o_p(1)$, and similarly,[8]

$$\frac{\mathbf{U}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{U}}{N} = \frac{\mathbf{b}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{b}}{N} + \frac{(\boldsymbol{\tau} \odot \boldsymbol{\sigma})' \mathbf{M}_{\boldsymbol{\Delta}} (\boldsymbol{\tau} \odot \boldsymbol{\sigma})}{N} + o_p(1) = \frac{\mathbf{b}' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{b}}{N} + \frac{\boldsymbol{\sigma}' \boldsymbol{\sigma}}{N} + o_p(1)$$

Consequently, $\hat{\beta}_{\text{OLS}}$ can be written as

$$\hat{\beta}_{\text{OLS}} = \left[1 - \frac{(\boldsymbol{\alpha} + \mathbf{b})' \mathbf{M}_{\boldsymbol{\Delta}} \mathbf{b} + \boldsymbol{\sigma}' \boldsymbol{\sigma}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}}\right] \beta + o_p(1)$$

which further suggests an adjusted version of $\hat{\beta}_{\text{OLS}}$, denoted by $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ below:

$$\hat{\beta}_{\text{OLS}}^{\text{adj}} = \frac{\hat{\beta}_{\text{OLS}}}{1 - \frac{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \hat{\mathbf{b}} + \hat{\boldsymbol{\sigma}}' \hat{\boldsymbol{\sigma}}}{\widehat{\boldsymbol{\alpha}}' \mathbf{M}_{\boldsymbol{\Delta}} \widehat{\boldsymbol{\alpha}}}}$$

[8] Here we use the independence of τ_i and τ_j, when $i \neq j$.

where $\hat{\mathbf{b}} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_N)'$, $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_N)'$, and $\hat{b}_i$ and $\hat{\sigma}_i$ are the approximated bias and standard error of $\widehat{\alpha}_i$, respectively.

5.2 Monte Carlo

5.2.1 Gini

The toy model we adopt for the Monte Carlo experiment is

$$y_i = \alpha_i + \epsilon_i,\ i = 1, 2, \ldots, N = 1{,}000.$$

where α_i stands for the population Gini coefficient for the ith group.

To mimic the small cross-sectional variation of Gini after control variables are projected out, the N values of α_i are specified in the following manner: α_i is equally distributed between 0.35 and 0.45, i.e., $\alpha_i = 0.35 + \frac{i-1}{10(N-1)}$. ϵ_i is drawn from a normal distribution with a zero mean and a variance equal to the sample variance of α_i. y_i is generated by the equation above.

For each α_i, the corresponding income data are randomly drawn from the Pareto distribution with the parameter $\lambda_i = (\alpha_i^{-1} + 1)/2$. Doing so is based on the fact that the c.d.f. of the Pareto distribution with parameter λ_i is $F(x) = 1 - x^{-\lambda_i}$, which implies that the population Gini coefficient is $1/(2\lambda_i - 1)$, according to the definition of Gini.

For convenience, the sample size of income data is fixed for every i, i.e., n is the number of observations used to compute $\widehat{\alpha}_i$, for every i. Note that choosing the same sample size in our Monte Carlo experiment does not imply that the sampling error of Gini is made homoscedastic: As i changes, both α_i and the income distribution function vary; consequently, the variance of sampling error is not fixed. The choice of n is listed in the first column of Table 5. With the simulated income data, we compute each sample Gini coefficient $\widehat{\alpha}_i$, the associated bias $\hat{b}_i$ and standard error $\hat{\sigma}_i$.

Finally, we compute three estimators of β, the parameter associated with Gini that is set to 1 in the toy model: (i) $\hat{\beta}_{\text{OLS}}$ is computed by regressing y_i on $\widehat{\alpha}_i$ with an intercept; (ii) $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ is the proposed adjusted version of $\hat{\beta}_{\text{OLS}}$ that accounts for the sampling error; (iii) $\hat{\beta}_{\text{EIV}}$ is the errors-in-variables estimator, which corresponds to $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ but with zero bias and the same standard error across groups.[9] With 1,000 replications, we report the bias and mean squared error (MSE) for the three estimators in Table 5.

As expected, Table 5 shows that $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ performs better than the unadjusted $\hat{\beta}_{\text{OLS}}$: both bias (in absolute value) and mean squared error of $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ appear substantially smaller than those of $\hat{\beta}_{\text{OLS}}$. As the sample size n increases, $\hat{\beta}_{\text{OLS}}$ is found to move toward β, as indicated by the decreasing values of bias and mean squared error. This is because the estimated sample Gini coefficient $\widehat{\alpha}_i$ becomes closer to the true population Gini coefficient α_i as the sample size increases. Similarly, the performance of $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ is also improved as the sample size increases. Overall, $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ consistently performs better than $\hat{\beta}_{\text{OLS}}$ as well as $\hat{\beta}_{\text{EIV}}$ that ignores the heterogenous feature of sampling error.

[9] We use $\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i$ as the standard error associated with Gini for $\hat{\beta}_{\text{EIV}}$.

Table 5 Bias and MSE of $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ by Monte Carlo–Gini

n	$\hat{\beta}_{\text{OLS}}$		$\hat{\beta}_{\text{OLS}}^{\text{adj}}$		$\hat{\beta}_{\text{EIV}}$	
	Bias	MSE	Bias	MSE	Bias	MSE
500	−0.773	0.599	0.189	0.112	−0.555	0.313
750	−0.724	0.524	0.139	0.055	−0.495	0.250
1,000	−0.684	0.469	0.107	0.037	−0.452	0.210
1,500	−0.623	0.390	0.078	0.019	−0.391	0.159
2,000	−0.575	0.332	0.057	0.013	−0.351	0.131
5,000	−0.421	0.181	0.033	0.005	−0.241	0.065
10,000	−0.311	0.100	0.021	0.003	−0.171	0.035

Bias $= \mathbb{E}(\hat{\beta} - \beta)$, MSE $= \mathbb{E}(\hat{\beta} - \beta)^2$. $\hat{\beta}_{\text{OLS}}$ ignores sampling error, $\hat{\beta}_{\text{OLS}}^{\text{adj}}$ accounts for sampling error, while $\hat{\beta}_{\text{EIV}}$ is the errors-in-variables estimator that (incorrectly) assumes sampling error has zero mean and the same standard error across groups. The reported numbers result from the average of 1,000 Monte Carlo replications. n is the number of observations used to estimate the unknown regressor α_i in $y_i = \alpha_i \cdot \beta + \epsilon_i$ where β is set to 1

Table 5 also highlights that β could be severely underestimated by $\hat{\beta}_{\text{OLS}}$, even when the sample size to compute the estimated regressor is large. For example, when the sample size is 10,000, the bias of $\hat{\beta}_{\text{OLS}}$ is found to be around −0.311. This implies that β is underestimated by roughly 31.1 % because β equals 1 in the Monte Carlo experiment.[10]

5.2.2 Sex ratio

We similarly conducted a Monte Carlo experiment for sex ratio, and it also supports our adjusted OLS estimator. We omit the detailed description of the experiment here, and codes are available by request.

References

Alesina A, Angeletos G-M (2005) Fairness and redistribution. Am Econ Rev 95(4):960–980
Angrist J (2002) How do sex ratios affect marriage and labor markets? Evidence from America's second generation. Quart J Econ 117(3):997–1038
Atkinson AB, Brandolini A (2001) Promise and pitfalls in the use of "secondary" data-sets: income inequality in OECD countries as a case study. J Econ Lit 39(3):771–799
Barro RJ (2000) Inequality and growth in a panel of countries. J Econ Growth 5(1):5–32
Barro RJ (2008) Inequality and growth revisited. Technical report. Asian Development Bank
Davidson R (2009) Reliable inference for the Gini index. J Econom 150(1):30–40

[10] In fact, having a cutoff for sample size, above which sampling error of Gini can be neglected is infeasible because (i) the sampling error of Gini depends on the sample size and the underlying income distribution; (ii) as indicated by the expression of $\hat{\beta}_{\text{OLS}}^{\text{adj}}$, the ratio of the sampling error and the variation of Gini determines whether sampling error can be neglected. Empirical researchers could, of course, compute this ratio to evaluate how severely the OLS estimator is contaminated.

De La Croix D, Doepke M (2003) Inequality and growth: why differential fertility matters. Am Econ Rev 93(4):1091–1113
Deininger K, Squire L (1996) A new data set measuring income inequality. World Bank Econ Rev 10(3):565–591
Deininger K, Squire L (1998) New ways of looking at old issues: inequality and growth. J Dev Econ 57(2):259–287
Deltas G (2003) The small-sample bias of the Gini coefficient: results and implications for empirical research. Rev Econ Stat 85(1):226–234
Edlund L, Li H, Yi J, Zhang J (2009) Sex ratios and crime: evidence from China. Rev Econ Stat 95(5):1520–1534
Forbes KJ (2000) A reassessment of the relationship between inequality and growth. Am Econ Rev 90(4):869–887
Hausman J (2001) Mismeasured variables in econometric analysis: problems from the right and problems from the left. J Econ Persp 15(4):57–67
Hoffman DL (1987) Two-step generalized least squares estimators in multi-equation. Rev Econ Stat 69(2):336–346
Jin Y, Li H, Wu B (2011) Income inequality, consumption, and social-status seeking. J Comp Econ 39(2):191–204
Kremer M, Chen DL (2002) Income distribution dynamics with endogenous fertility. J Econ Growth 7(3):227–258
Langel M, Tillé Y (2013) Variance estimation of the Gini index: revisiting a result several times published. J R Stat Soc Ser A (Stat Soc) 176(2):521–540
Lorenz MO (1905) Methods of measuring the concentration of wealth. Publ Am Stat Assoc 9(70):209–219
Modarres R, Gastwirth JL (2006) A cautionary note on estimating the standard error of the Gini index of inequality. Oxf Bull Econ Stat 68(3):385–390
Murphy KM, Topel RH (1985) Estimation and inference in two-step econometric models. J Bus Econ Stat 3(4):370–379
Ogwang T (2000) A convenient method of computing the Gini index and its standard error. Oxf Bull Econ Stat 62(1):123–129
Sandström A, Wretman JH, Walden B (1988) Variance estimators of the Gini coefficient—probability sampling. J Bus Econ Stat 6(1):113–119
Sen A (1973) On economic inequality. Oxford University Press, Oxford
Stock JH, Wright JH, Yogo M (2002) A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat 20(4):518–529
Wei S-J, Zhang X (2011) The competitive saving motive: evidence from rising sex ratios and savings rates in China. J Polit Econ 119(3):511–564
White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4):817–838