ECINEQ
Society for the Study
of Economic Inequality

Working Paper Series

The rich underreport their income: Assessing bias in inequality estimates and correction methods using linked survey and tax data

Sean Higgins
Nora Lustig
Andrea Vigorito

ECINEQ WP 2018 - 475

ECINEQ 2018 - 475
September 2018

www.ecineq.org

The rich underreport their income: Assessing bias in inequality estimates and correction methods using linked survey and tax data*

Sean Higgins
UC Berkeley, USA

Nora Lustig
Tulane University, USA

Andrea Vigorito
Instituto de Economía, FCEA, Universidad de la República, Uruguay

Abstract

Do survey respondents misreport their income? If so, how does misreporting correlate with income, how does this affect estimates of income inequality, and how well do existing methods correct for bias? We use a novel database in which a subsample of Uruguay's official household survey has been linked to tax records to document the extent and distribution of labor income underreporting and to assess the performance of various existing methods to correct inequality estimates. Individuals in the upper half of the income distribution tend to report less labor income in household surveys than those same individuals earn according to tax returns, and underreporting is increasing in income. Using simulations, we find that this leads to downward-biased inequality estimates. Correction methods that rely only on survey data barely affect the biased inequality estimates, while methods that combine survey and tax data can lead to over-correction and overestimation of inequality.

Keywords: inequality, income underreporting, tax records, household surveys.

JEL Classification: D31, C81.

*We thank Uruguay's Instituto Nacional de Estadística and Dirección General Impositiva for providing the data used in this research. We are grateful to François Bourguignon and seminar participants for providing useful comments, and to Mauricio De Rosa and Siyu Quan for research assistance. This study was made possible thanks to the generous support of the Bill & Melinda Gates Foundation (grant number OPP1135502) and Uruguay's Universidad de la República.

1 Introduction

In recent years, income inequality has increasingly captured public attention. Newspaper headlines highlight the exorbitant income share of the top 1%, or compare CEO pay to ordinary workers' wages. Household surveys do not capture incomes at the top of the income distribution well, however (Atkinson, 2007): the rich may be harder to reach (e.g., in gated communities), more likely to refuse to answer the survey when reached, or may report a lower fraction of their income when responding to the survey.

Recognizing this, recent papers have made use of data from tax returns to correct survey-based inequality estimates. A variety of methods have been developed, but each one relies on implicit assumptions that are often untestable because there is no way to link survey income to the same individual's income according to her tax returns. In this paper, we exploit a novel data set that directly links a subset of individuals from Uruguay's official household survey to the same individuals' tax returns, enabling us to observe labor income from each of these sources for the same person.

We make two main contributions. First, we document the extent and distribution of labor income underreporting in surveys. Second, we assess the performance of various existing methods to correct inequality estimates, which had not previously been testable without information from linked survey and tax data. We find that individuals in the upper half of the income distribution tend to report less labor income in household surveys than those same individuals earn according to tax returns, and underreporting is increasing in income. Using simulations, we find that this leads to downward-biased inequality estimates. Correction methods that rely only on survey data barely affect the biased inequality estimates, while methods that combine survey and tax data can lead to nontrivial overestimation of inequality.

2 Methods to correct inequality estimates

The main causes for the lack of accurate information on top incomes in household surveys are frame or noncoverage errors, unit and item nonresponse, underreporting, and top coding (Biemer and Christ, 2008; Cowell and Flachaire, 2015). These issues can lead to significant bias in inequality measures, and this bias can be either positive or negative (Deaton, 2005).

As summarized in Lustig (2018), the approaches to correct for the missing rich in household surveys can be distinguished by whether the method relies on the income data reported in the survey alone, or combines these data with external sources such as tax records and/or national accounts. An example of the within-survey methods is replacing incomes above percentile p with a Pareto (or other parametric) distribution fit to the survey data above p (Cowell and Flachaire, 2007). Within-survey replacing of top incomes by a parametric distribution should have a limited effect on the inequality estimate because the corrected distribution is still fit to observations from the survey. Another option is to incorporate tax data and replace incomes above p in the survey by a Pareto Type I or II distribution fit to *tax data* rather than survey data (Atkinson, 2007; Jenkins, 2017).

Another key distinction is whether the method replaces the incomes of the observations in the top tail or reweights population shares. The first approach assumes that the population shares of top incomes (the rich) and the rest (the nonrich) in the survey are correct, but that incomes captured at the top are incorrect or incomplete (due, for example, to underreporting). As discussed in Cowell and Flachaire (2015), a major challenge in this approach is to find the threshold above which the correction should be applied. Examples of this *replacing* approach include the parametric approaches of fitting a distribution above p discussed above. Another is the semiparametric approach from Piketty et al. (2017), who calculate the ratio of tax income to survey income at each percentile and scale up income above p by that percentile's ratio, then scale up incomes between some $p' < p$ and p by a factor that rises linearly from 1 (at p') to the observed ratio at p.[1]

The second correction approach, *reweighting*, assumes that the population weights for the rich and nonrich in the sample are incorrect: one must "add people" at the top by giving the observations in the right-hand tail a higher weight, or even by adding new observations and compressing

[1] More precisely, Piketty et al. (2017) use percentiles below $p = 0.99$ but then use increasingly fine "generalized percentiles" within the top 1% of the income distribution. We also use these generalized percentiles when we implement their correction method in Section 4.

the weights of observations in the survey. The latter case, used by Anand and Segal (2017), implicitly assumes that the observations in the survey do not represent the target population due to noncoverage or unit nonresponse.

3 Data

We use a novel database in which a subsample of Uruguay's official household survey—the *Encuesta Continua de Hogares* (ECH)—has been linked to personal income tax records from the *Dirección General Impositiva* (DGI). We did not have access to the actual card numbers (*cedula*), but to masked identifiers. The DGI databases include the universe of potential personal income tax payers, including all formal workers, for 2009–2014 (Burdín et al., 2014). Like all tax records, these data are subject to tax evasion and avoidance.

The subsample of the 2012 and 2013 ECH that were linked to tax records belong to a follow-up survey, the Nutrition, Child Development, and Health Survey (*Encuesta de Nutrición, Desarrollo Infantil, y Salud*, ENDIS), which includes 2,709 urban households with children aged 0 to 3. ENDIS collected the national identification number of each respondent, which allowed us to merge mothers from ECH/ENDIS with DGI tax records. Thus, we can compare tax-return incomes to the same individuals' survey incomes, but a limitation of our analysis is that it is restricted to mothers with children aged 0 to 3 living in urban areas.

Of the 2,704 ENDIS respondents, 1,412 declared being employed; of these, 1,022 were matched to the 2009–2014 DGI database. We restrict our analysis to the 807 matched households with positive labor earnings in the DGI records for the month preceding the ECH interview.

We restrict our analysis to formal labor income.[2] The income concept from tax records is post-tax formal labor earnings for the month prior to the ECH interview.[3] In ECH, we construct a measure of post-tax formal labor income defined identically to our income measure from tax records by adding salaries and wages, commissions, incentives, overtime payments, tips, arrears, transport, food or housing vouchers, other in-kind payments, other fringe benefits, bonuses, and

[2] Informal labor income is excluded in our income measures from both sources because it is not captured in the tax-return data.

[3] These incomes are reported on forms 1444, 3100, 1102, 1103, 1104, and 1201.

vacation pay, for each formal occupation held by the respondent.

4 Results

4.1 Comparing labor income from surveys and tax returns

We use the linked data to examine income misreporting throughout the distribution. Figure 1 plots the ratio of income reported in the survey to income from tax returns for each observation in the linked data, and shows how this varies across the tax return distribution (i.e., by tax return income percentile). If everyone reported the same income in the two data sources, all points would lie along the $y = 1$ line, shown in orange. If incomes were reported with noise but income misreporting were orthogonal to income, the points would bounce around, but the average relationship would correspond to the orange line at $y = 1$. Survey incomes exceed tax return incomes in approximately the bottom half of the tax return distribution, while survey incomes are lower in the top half (as seen by the blue line, which shows a local polynomial regression of the points in the scatterplot). Looking at averages by income percentile (shown by larger, green dots), the top 1% of the tax return distribution reports only about 60% of the labor income from their tax returns in the household survey.

Figure 2 shows the empirical copula (i.e., bivariate density) of percentiles in the survey and tax return income distributions. If the correlation between every individual's rank in the tax return income distribution and her rank in the survey income distribution were the same (which can occur regardless of the extent of misreporting), all of the copula's density would lie along the gray 45-degree line. We see that at the lower end of the distribution, there is a lot of noise, and being in a certain percentile of one distribution has less predictive power about where an individual's income falls in the other distribution. Furthermore, those at low percentiles in the survey-based income distribution tend to be found at higher percentiles in the tax return distribution, as seen by the solid blue line which graphs a local polynomial regression of the correlation in ranks. At higher incomes, however, the correlation is stronger: among the top 20% of the income distribution we see a higher density of observations concentrated near the 45-degree line, although those in the highest survey income percentiles tend to be found at slightly lower tax-return income percentiles due to misreporting.

Finally, because some of the methods to correct inequality estimates assume the same support between the household survey and tax return income distributions, we test whether—again restricting ourselves to the linked sample and thus ruling out some reasons for differing supports (such as the rich being less likely to respond to the survey, or the survey being a sample)—the two income distributions have different supports. Figure 3a shows histograms of the income distributions above the 90th-percentile survey income as in Jenkins (2017). We see that the supports are not the same: even though we have restricted the analysis to the same individuals in the two data sources, the maximum income from tax records is about twice as large as the maximum income from the survey. We can view this more clearly among the top 1% from the survey: Figure 3b zooms into incomes above the 99th-percentile survey income, where we see substantially higher mass in the tax returns income distribution.

4.2 Applying correction methods to survey data

We now test the methods to correct inequality estimates described in Section 2. We test these using a hypothetical true income distribution based on the DGI universe of tax returns from 2012–2013—restricted to tax filers with income greater than or equal to the monthly minimum wage—then simulate misreporting based on the patterns observed in our linked data set.[4] We use simulation for two reasons. First, tax data might also be misreported, so we cannot merely apply the correction methods to survey data and compare the result to our inequality estimate from tax data; instead we begin with a hypothetical true distribution and simulate misreporting on it. Second, because the linked sample is small and represents one of many possible realizations from a survey, simulations allow us to estimate the effect of correction methods under many realizations and with a larger sample.

Our simulation method is as follows. We first estimate a smoothed measure of average mis-

[4] We restrict to incomes of at least the monthly minimum wage since the tax data are less reliable below that threshold according to DGI, so differences between survey and tax data below that threshold might not be due to misreporting in the survey.

reporting $\hat{\mu}_p$ in each tax return percentile p of the linked data using a local linear regression; this corresponds to the solid blue line in Figure 1. Within each percentile of the linked data we also estimate a smoothed standard deviation of misreporting $\hat{\sigma}_p$; intuitively, this corresponds to the spread of the gray dots in Figure 1 for a given percentile on the x axis. We then turn to the 2012–2013 tax data from the universe of potential tax filers, compute percentiles prior to restricting the data, then restrict the data to those reporting at least the monthly minimum wage (which restricts the data to the top 61% of incomes). For each observation i, we sample a misreporting factor f_i from a normal distribution, $f_i \sim N(\hat{\mu}_{p(i)}, \hat{\sigma}_{p(i)})$. We multiply i's tax return income by f_i to obtain her misreported income.

We then apply the correction methods to each simulated misreported income distribution. We test 6 correction methods: within-survey replacing with a Pareto I and II above $p = 0.9$ (Cowell and Flachaire, 2007); replacing using a Pareto I and II fit to *tax data* above $p = 0.9$ (Atkinson, 2007; Jenkins, 2017); Piketty et al.'s (2017) replacing method using tax data; and Anand and Segal's (2017) method combining reweighting, tax data, and finally a Pareto I above $p = 0.9$. Whenever a Pareto is fit to the top of the income distribution, we use Alvaredo's (2011) formula to estimate the Gini coefficient of inequality.

Figure 4 shows the results. While the true Gini is 0.386, in our simulations of misreporting, the estimated Ginis range from 0.369 to 0.373 with a mean of 0.371 (Figure 4a).[5] The within-survey replacing methods hardly change the (biased) estimate of inequality from the misreported income distribution, and sometimes lead to even lower inequality estimates: the mean Gini estimates after within-survey Pareto I and II corrections are 0.369 and 0.370. Using tax data to fit parametric distributions partially corrects the bias, with mean Gini estimates of 0.377 (for both Pareto I and II). Replacing using tax data above percentile 99.5 and scaling up survey incomes between percentiles 80 and 99.5 (Piketty et al., 2017) leads to over-correction, as does the reweighting method from Anand and Segal (2017): these methods lead to average Gini estimates of 0.412 and 0.406,

[5] The top 10% income share ranges from 0.286 to 0.290 with a mean of 0.288, compared to 0.314 in the true distribution (Figure 4b).

respectively.[6] This over-correction is not specific to the Gini, and also occurs when we use other inequality measures such as the top 10% income share (Figure 4b).

5 Conclusion

We use linked household survey-tax return data to isolate one of the problems plaguing inequality measurement based on household surveys: income underreporting (particularly among the rich). We document that individuals in the upper half of the labor income distribution tend to report less income in household surveys than those same individuals earn according to tax returns, and that underreporting is increasing in income. The top 1% reports, on average, about 60% of their labor income (according to tax returns) in the household survey.

Within-survey replacing methods barely affect the inequality estimate (i.e., it remains biased). Turning to correction methods that combine survey and tax data, we find that both replacing and reweighting can lead to over-correction and overestimation of inequality. As a result, studies that correct survey data using a single correction method to assess the level and trend of inequality—as well as studies that make no correction and rely only on incomes reported in surveys—should be interpreted with caution.

[6] While the Anand and Segal (2017) method implicitly assumes that none of the top 1% of the true income distribution is captured in the survey, here we make an assumption at the opposite extreme by only focusing on misreporting. Nevertheless, the richest observation in our linked data belongs to the top 0.1% of the full tax returns distribution, which is evidence against the assumption that none of the top 1% are captured in the survey.

Figure 1: Ratio of survey income to tax return income

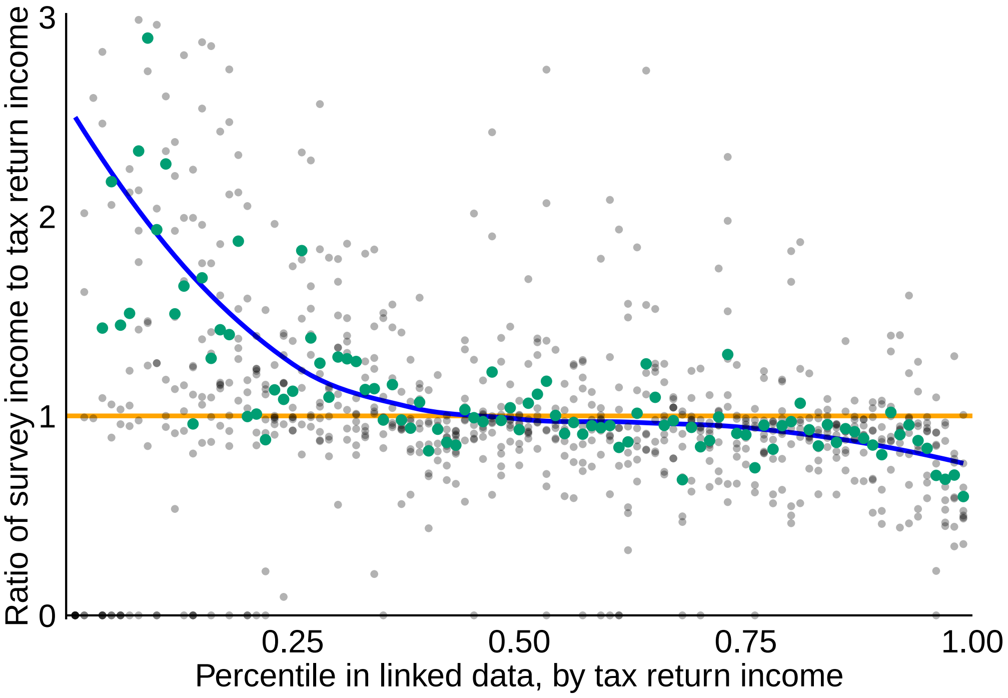


Figure 2: Bivariate distribution of rank in tax return and household survey distributions

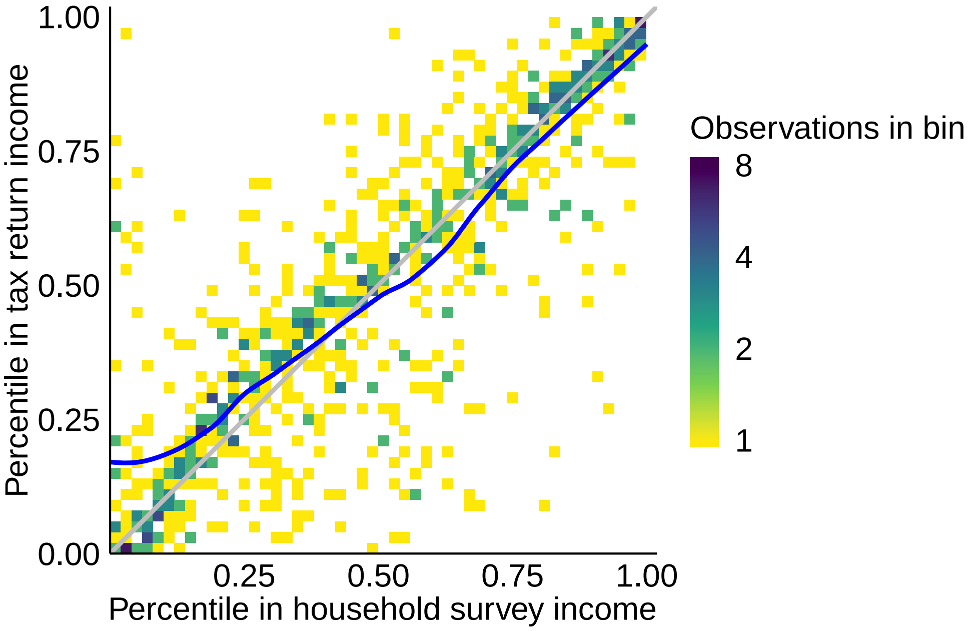
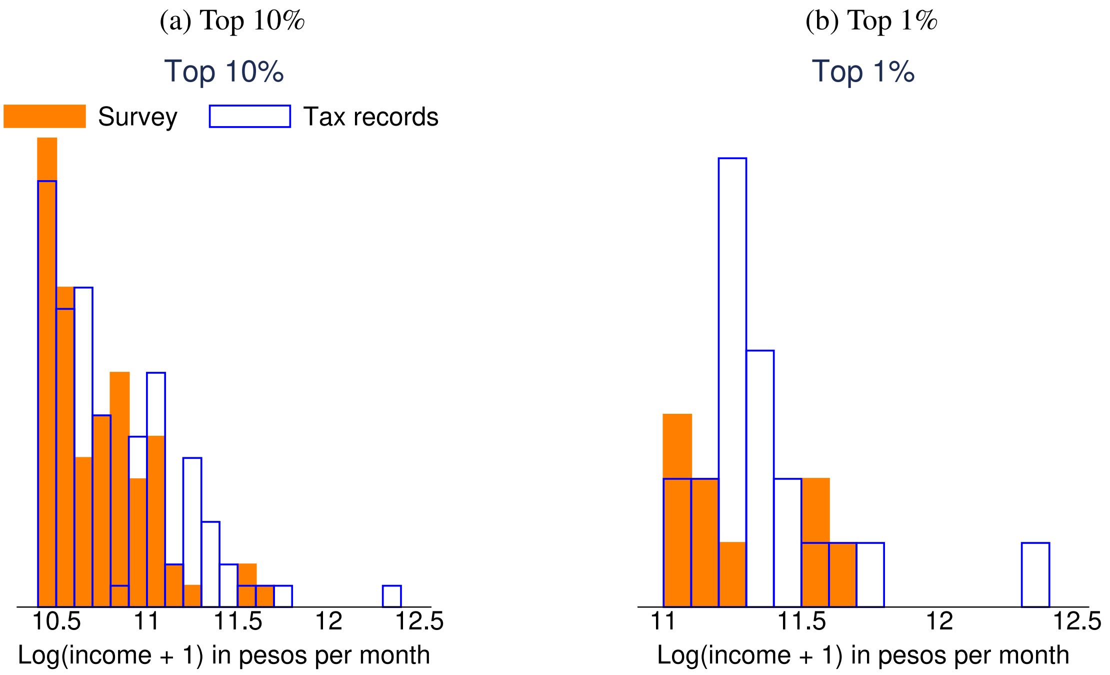


Figure 3: Zoomed-in anonymous income distributions from survey and tax records, linked sample

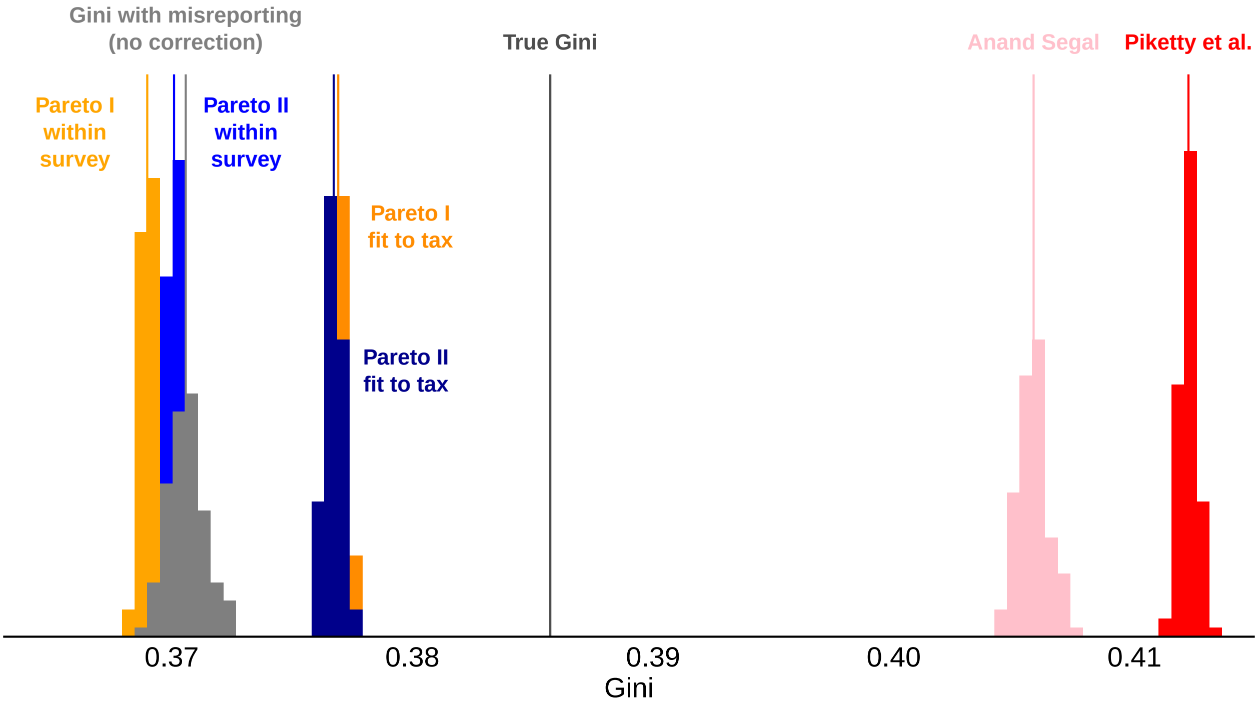


(a) Top 10%

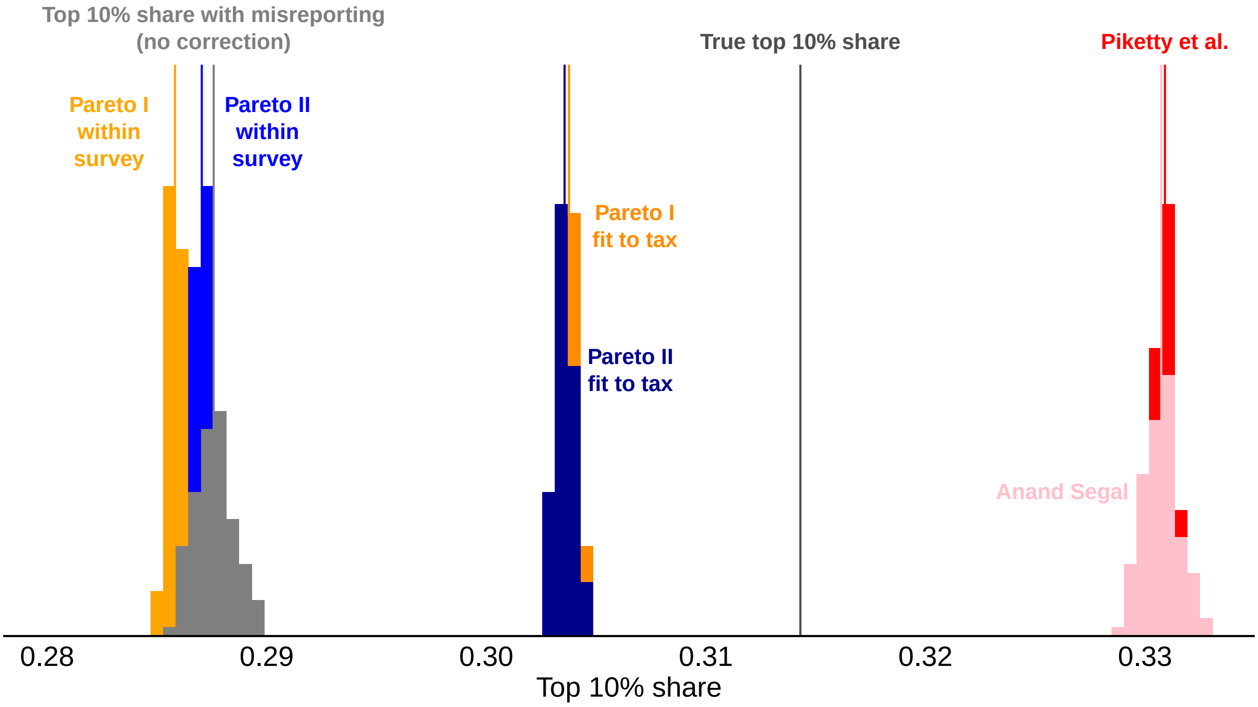
(b) Top 1%

Figure 4: Inequality estimates with misreporting, under various correction methods

(a) Gini coefficient



(b) Top 10% income share



References

Alvaredo, F. (2011). A note on the relationship between top income shares and the Gini coefficient. *Economics Letters 110*, 274–277.

Anand, S. and P. Segal (2017). Who are the global top 1%? *World Development 95*, 111–126.

Atkinson, A. B. (2007). Measuring top incomes: methodological issues. Volume 1. Oxford University Press New York, pp. 18–42.

Biemer, P. P. and S. L. Christ (2008). Weighting survey data. In *International Handbook of Survey Methodology*. pp. 317–341.

Burdín, G., F. Esponda, and A. Vigorito (2014). Inequality and top incomes in Uruguay: a comparison between household surveys and income tax micro-data. Commitment to Equity Working Paper 21.

Cowell, F. A. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics 141*, 1044–1072.

Cowell, F. A. and E. Flachaire (2015). Statistical methods for distributional analysis. In *Handbook of income distribution*, Volume 2. Elsevier, pp. 359–465.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and Statistics 87*, 1–19.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in uk income inequality. *Economica 84*(334), 261–289.

Lustig, N. (2018). The missing rich in household surveys: Causes and correction approaches. CEQ Working Paper 75.

Piketty, T., L. Yang, and G. Zucman (2017). Capital accumulation, private property and rising inequality in China 1978-2015. NBER Working Paper 23368.