Top Incomes, Issues with Survey Data, and Inequality: Evidence from Simulations and Linked Income and Tax Return Data[1]

Sean Higgins (UC Berkeley)
Nora Lustig (Tulane University)
Andrea Vigorito (Universidad de la República)

LACEA–LAMES
Buenos Aires
November 9, 2017

[1] This research project was conducted for the Commitment to Equity (CEQ) Institute. The study was made possible thanks to the generous support of the Bill & Melinda Gates Foundation. For more details about the CEQ Institute, visit www.commitmentoequity.org.

Motivation

- Many issues with survey data
  - Lead to biased inequality estimates (Chesher and Schluter, 2002; Cowell and Flachaire, 2007)
- Often these issues lead to the "missing rich"
  - Underestimation of income at the top
  - Resulting bias in inequality estimate can be substantial!
- Various corrections proposed (Atkinson, 2007; Jenkins, 2017; Campos-Vazquez and Lustig, 2017)
  - These corrections make a lot of assumptions
  - Mostly untestable (until now)

This paper

- Novel data set on linked household survey data and tax returns from Uruguay
- Assuming tax return data is “correct” (for now):
  - Examine misreporting of labor income in household survey
  - Examine undercoverage
    - Biases in survey design (sampling frame not the same as target population)
    - Unit non-response (individuals cannot be reached or refuse to respond)
- Simulate these issues on full tax return data set
- Simulate proposed corrections and see how well they work

Data: Tax Returns

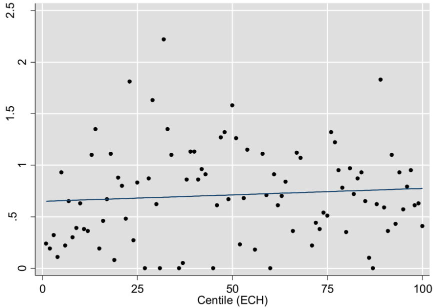
- Universe of potential tax payers, 2009–2014
- $\sim$1.9 million observations per year
- Main variables:
  - Pre- and post-tax annual income by source
  - Monthly labor earnings
  - Taxes
  - Deductions
  - Sex, age, industry, firm characteristics
- Around 33% of workers in data are above minimum threshold and thus pay taxes
- Limitations: evasion, avoidance, non-taxable rents

Data: Household Survey

- Encuesta Continua de Hogares
  - Income and labor force status from 2012–13 wave interviews
  - We focus on labor income
  - Nationally representative sample
  - Sample size: 46,550 in 2012 and 46,669 n 2013
- Follow-up nutrition survey on subsample ($N = 2704$)
  - This is the survey that asked identifiers to merge with tax data
  - Mothers with children aged 0–3
  - For now we can only use this subsample
  - But working with statistical institute to do analysis on full survey sample in ongoing wave

Merged Data

- Of the 2704 in ECH follow-up survey:
  - 1236 merged (1412 in ECH follow-up declared being employed)
  - 775 with positive labor earnings in month preceding ECH interview
- This is our final sample for merged data tests



Simulations

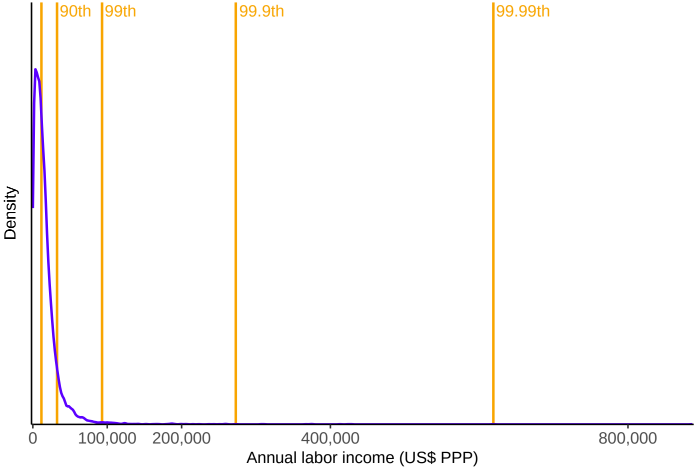
- We simulate three types of issues, using linked Uruguay data to guide functional form of simulations:
  1. Misreporting
     - People respond to survey but report incorrect income amounts; possibly correlated with income
  2. Undercoverage
     - Unit non-response: some do not respond to survey
     - Bias: sampling frame not the same as target population
     - Possibly correlated with income
  3. Extreme observations
     - Everyone sampled responds to survey, but what is effect of only sampling some of top incomes?

Corrections

- We simulate two types of corrections supposing we have some information about the "true" income distribution
  1. Reweighting
     - à la Campos-Vazquez and Lustig (2017)
     - Suppose you know density of people within each of a number of income bins
     - Reweight in survey to match that density
  2. Adjusting incomes
     - Suppose you know mean income by group (e.g. decile) of "true" distribution
     - Scale incomes in the survey within each group to match incomes in true distribution
- Identical in theory (Bourguignon, 2017)
  - If continuous distribution with same support

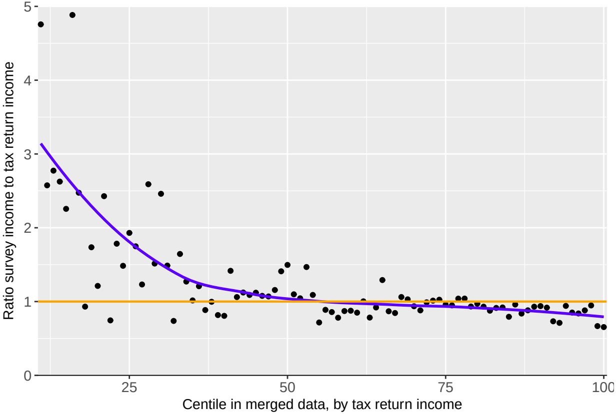
Distribution for Simulations

- Use full distribution of positive labor income earners ($N = 1.3$ million) from Uruguay tax returns data

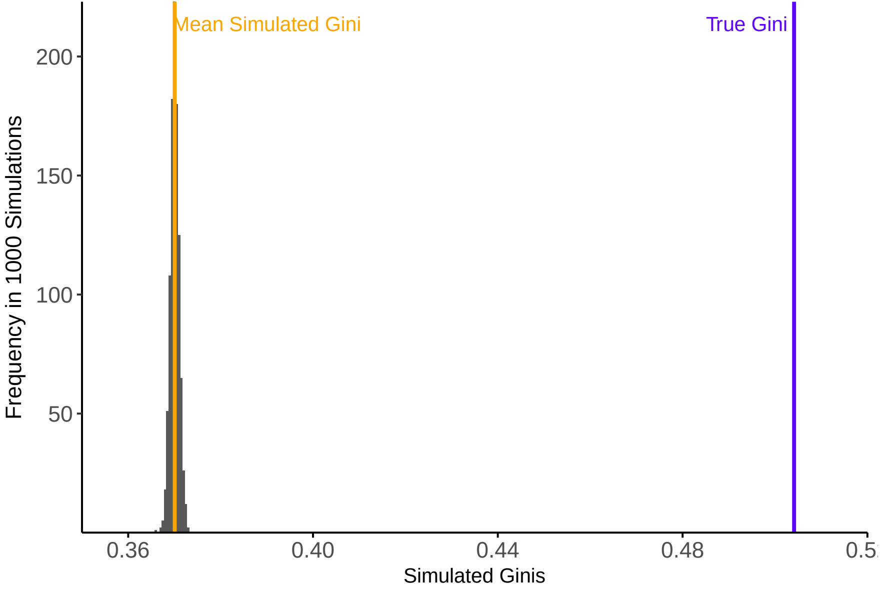


Misreporting

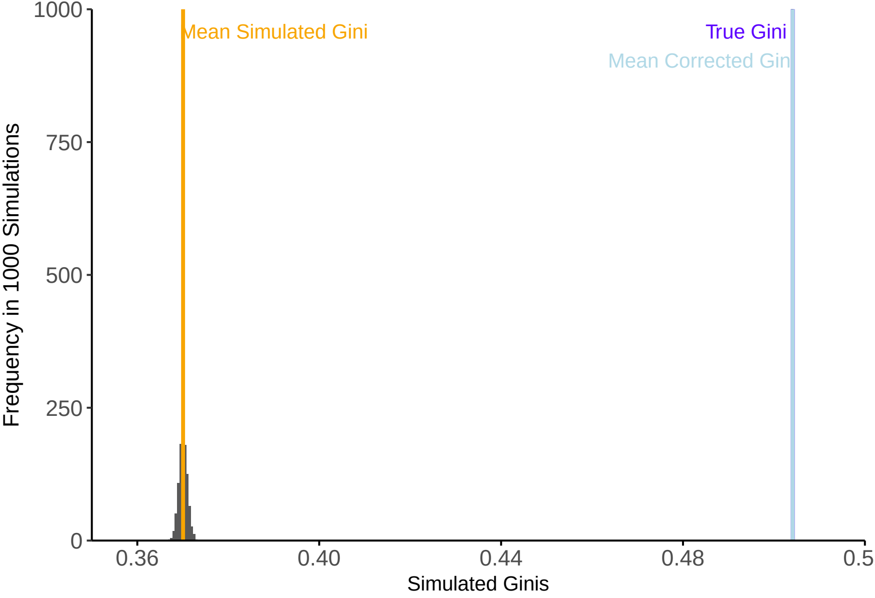
- We impose the same relationship between income and misreporting as that observed in the merged Uruguay data
  - Using a non-linear Loess regression



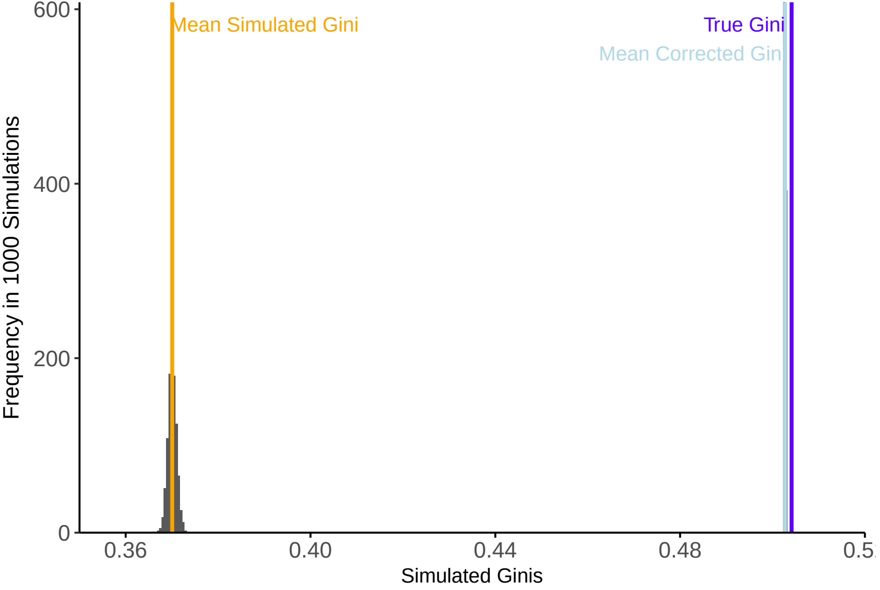
Misreporting



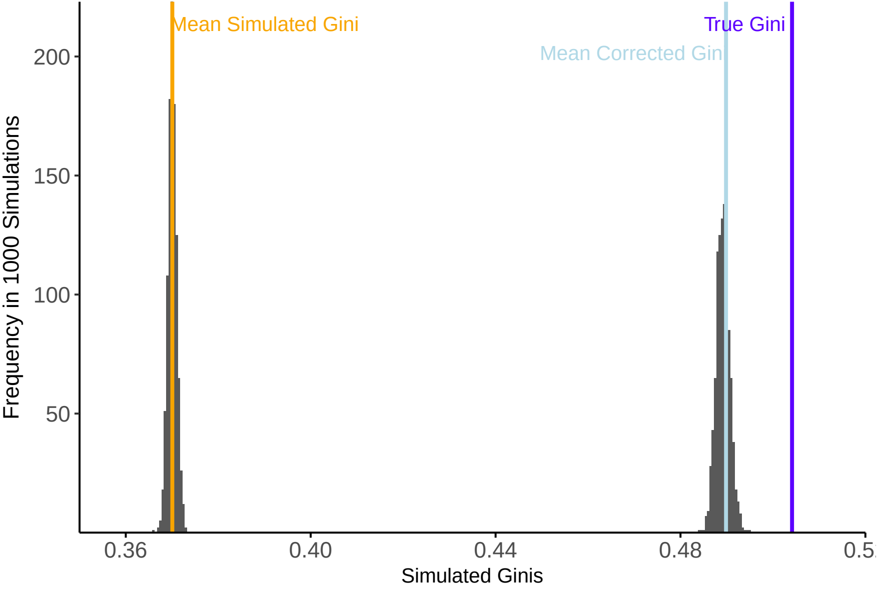
Misreporting, Income Adjustment by Centile



Misreporting, Income Adjustment by Decile

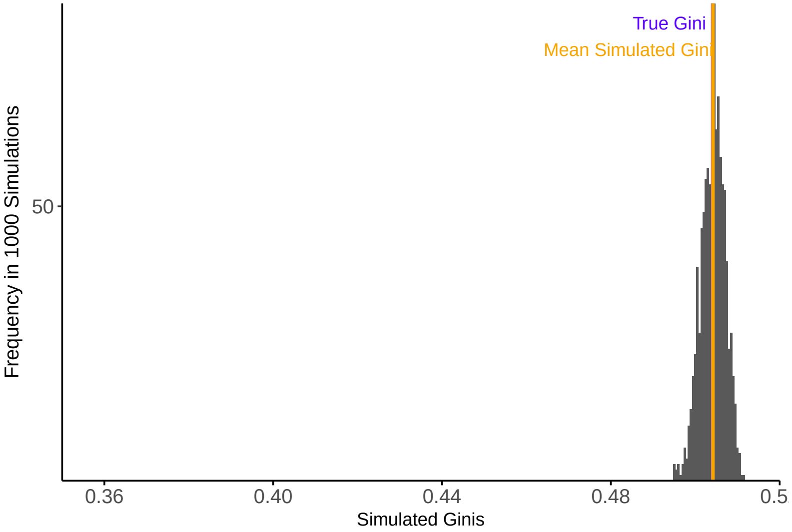


Misreporting, Reweighting Correction



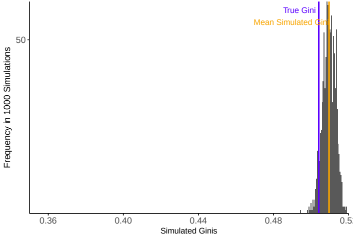
Random Non-response

- Assume $P(n) = .2$, independent of income



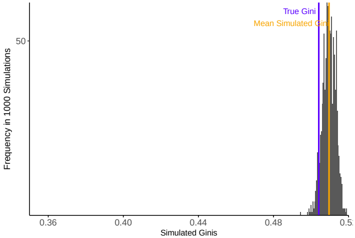
Non-response Increases with Income

(a) $P(n) = .1 + .2F(y)$

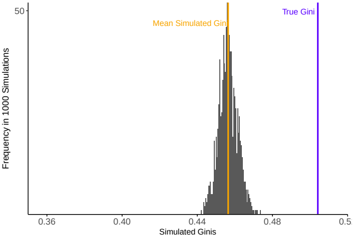


Non-response Increases with Income

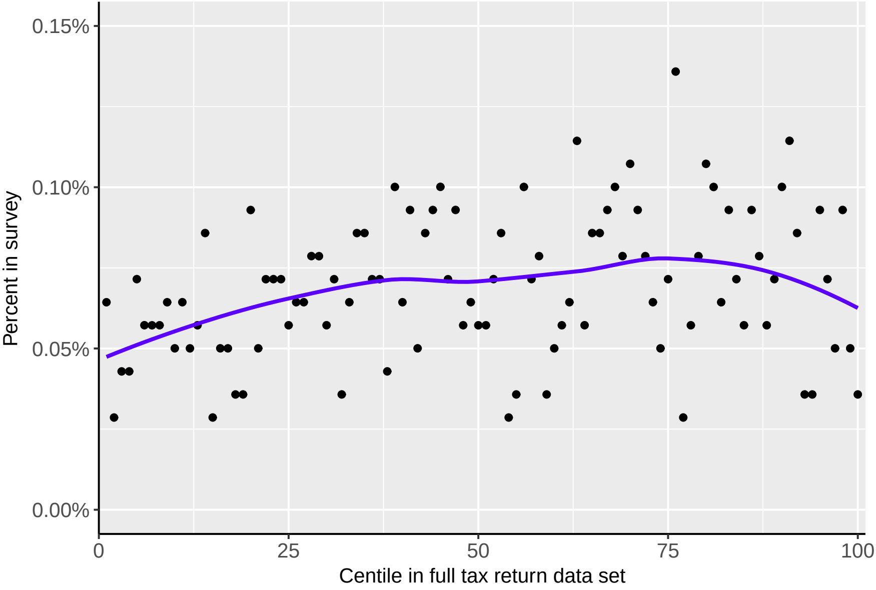
(a) $P(n) = .1 + .2F(y)$



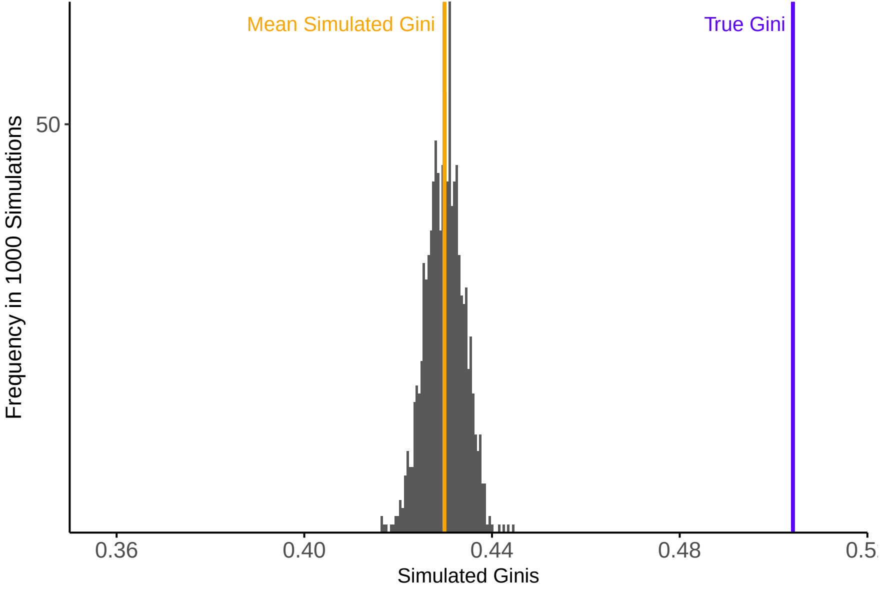
(b) $P(n) = \mathbb{I}(F(y) > .7) \cdot (.567 + \frac{2}{3}(F(y) - .7))$



Non-response as in Uruguay



Non-response as in Uruguay



Extreme Observations

- 1% sample within each percentile