DataFirst

DataFirst Technical Papers

Problems with SWIID: the case of South Africa

by
Martin Wittenberg

Technical Paper Series
Number 30

About the Author(s) and Acknowledgments

Martin Wittenberg - Director of DataFirst and Professor in the School of Economics, University of Cape Town
This is a joint SALDRU Working Paper and DataFirst Technical Paper.

Recommended citation

Wittenberg, M. (2015). Problems with SWIID: the case of South Africa. A DataFirst Technical Paper 30. Cape Town: DataFirst, University of Cape Town

---

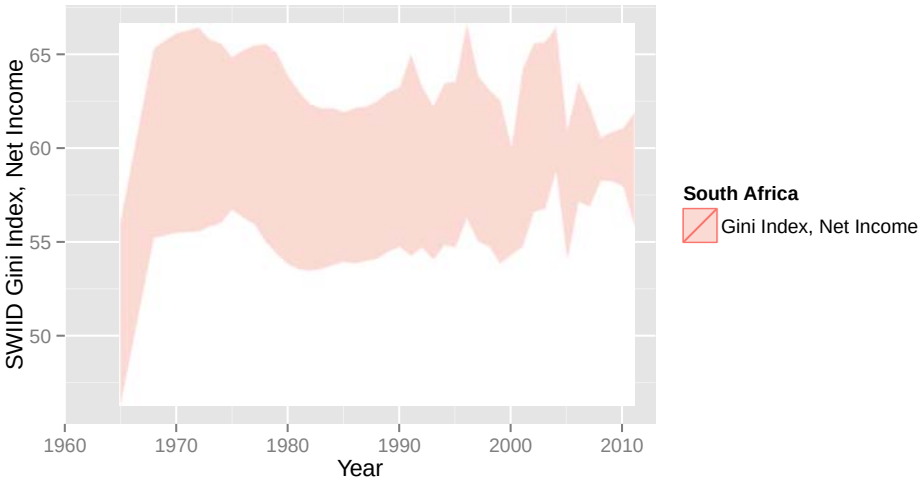
DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

Problems with SWIID: the case of South Africa

Martin Wittenberg

DataFirst Technical Paper 30
University of Cape Town
June 2015

The information contained in databases of summary statistics should look plausible when viewed in context. Judged by that criterion the Standardized World Income Inequality Database (SWIID) comes up short with its South African data. Figure 1 contains the series as extracted from the SWIID web site (Solt 2014b). The 95% confidence bands suggest that inequality in 1965, at the height of apartheid, was significantly lower than in the 2000s. This, however, flies in the face of much other evidence. For instance it is well-known that Black mine workers' wages were static in real terms from the early twentieth century right up to the 1970s (e.g. van der Berg 1989). The wages of white miners, by contrast, increased, so that the ratio of White to Black mine wages reached its maximum of twenty to one in 1969 (Devereux 1983, p.18). Simkins (1979) estimated the Gini coefficient in 1970 at 0.71, which seems more in line with the political and social realities. Trying to understand how the SWIID may have arrived at such a misleading estimate[1] is instructive about the types of problems that may be lurking elsewhere in the database.



Note: Solid lines indicate mean estimates; shaded regions indicate the associated 95% confidence inter vals.
Source: Standardized World Income Inequality Database v5.0 (Solt 2014).

There are four potential sources of error in SWIID: measurement error, model error, imputation error and sampling error[2]. We will discuss these in turn.

[1] The number in version 4 of the SWIID was even more out of line. The point estimate (i.e. average of the replicates) was 0.43, compared to 0.51 in version 5 of the database.

2 For a more detailed review of the SWIID see Jenkins (forthcoming).

1. Measurement error

Coverage problems (e.g. only surveying urban populations), problems with the census or survey instrument (e.g. omitting to measure certain types of income) or response issues can all lead to seriously biased estimates. Bad measurements can, in turn, affect other estimates through the standardization protocols of the SWIID. Some care therefore needs to be taken that the initial data used in the SWIID algorithms is fit for purpose. Solt is alive to these difficulties. For instance, he omits some data points for Russia because they "seem to lack face validity" (Solt 2014a, p.7). There is no explicit discussion of the South African data, but the replication code indicates that the Gini coefficients reported by Lachman and Bercuson (1992) have been excluded. The remaining estimates for the 1960s seem all to be taken from the World Income Inequality Database (WIID) version 2c. This has three Gini coefficients for 1965 due to Jain (1975), Lecaillon *et al*. (1984) and Paukert (1973). McGrath has commented on the Jain estimate as being "patently deficient to anyone who has the slightest knowledge of the demography and geography of the economy" (1984, p.3). Indeed **all** the estimates are deficient because they do not adequately deal with the fact that the majority of the South African population, which also happens to be the poorest section, is largely missing from the statistics (e.g. McGrath 1984, p.58). As Devereux (1983) explains:

> "until the 1980 census, government censuses and surveys specifically ignored the personal incomes of Blacks." (p.2) …
> [Consequently] "Black incomes must be calculated as a residual of Total Personal Income" (p.5),

i.e. the incomes reported in the censuses (for Whites, Coloureds and Indians) are compared to national accounts estimates to obtain some idea what Black South Africans might have been paid. Devereux also notes that academics tried to get around this data deficiency by using disparate sources that happened to be available (p.2). Simkins (1979), for example, combines census information (for Whites, Indians and Coloureds) with income and expenditure surveys conducted by the Bureau for Market Research among Black South Africans in selected urban areas. The estimates of the Gini coefficients obtained in this way are markedly higher than the ones contained in the WIID (see UNU-WIDER 2014). Nevertheless even the defective estimates in the WIID range between 0.56 and 0.58 which is higher than the SWIID point estimate shown in Figure 1. Clearly the standardization also plays a part.

2. Model Error

The next step in the SWIID algorithm is to convert the "raw" Gini coefficients from the source data to standardized ones, using ratios estimated for this purpose. The types of conversion are between different welfare definitions, i.e. net income, market income and expenditure (Solt 2014, p.8) and different equivalence scales, viz. "(1) household per capita, (2) household adult equivalent, (3) household unadjusted, or (4) person" (Solt 2014, p.8). This gives eleven types of conversion, given that person level information is never combined with expenditure data.

Like with any estimation process, the quality of the predicted values emanating from this procedure depends on whether the model describes the underlying process correctly. In this case we are assuming that the conversion ratios are constant within regional groupings. Is this a good assumption? In the case of South Africa it would mean that it is being compared to the likes of Egypt,

Gabon, or Kenya (all of which have data from the 1960s[3]) or to South Africa post-apartheid. The key question is whether one thinks of inequality in apartheid era South Africa as being *sui generis* or comparable to that in postcolonial countries. Assuming constancy may be the best that one can do, in the absence of better information, but it is clear that if the ratios **were** different, then the "standardized" coefficient may be off, even if the raw data were measured without bias.

There is another step worth noting that may introduce error. The coefficients are smoothed by a moving average process and in the process coefficients for intervening years are interpolated. The smoothing prevents abrupt changes, which may be reasonable for most "normal" countries, but it may not be so valid for countries undergoing major crises. Indeed Solt dispenses with the smoothing for Eastern Europe after the demise of communism.

3. Imputation Error

Even if the data are measured correctly and the imputation model is correct, the standardized coefficients are still predictions and will not be exactly equal to the true Gini coefficient (for that welfare concept and equivalence measure), if that had been measured. Solt deals with this source of uncertainty by presenting not the mean prediction, but one hundred draws from the distribution of possible outcomes (assuming the errors are normally distributed), given the imputation model[4].

4. Sampling Error

There is a final source of error, even for the "gold standard" measurements as contained in the LIS. Survey based measures of the Gini coefficient are subject to sampling error, given that a different sample would have obtained slightly different coefficients. This type of error is also accounted for in the way the replicates are drawn.

Lessons from the South African case

The 95% confidence intervals as shown in Figure 1 deal with the third and fourth type of error only. They do not correct for bad measurements, inappropriate conversions and erroneous interpolations. Bad measurements, of course, are a problem for other databases too – using the existing WIID figures for South Africa for the 1960s would also be problematic. But the promise in the SWIID is that it deals with the "noise" in the data:

> "In version 5.0 of the SWIID, the inequality estimates and their associated uncertainty are represented by 100 separate imputations of the complete series: for any given observation, the differences across these imputations capture the uncertainty in the estimate." (Solt 2014c, p.1)

In the case of South Africa in 1965 even the full range of the 100 imputations does not get close to what the correct Gini is likely to have been.

---

[3] The others with 1960s data in SWIID are Madagascar, Morocco, Niger, Senegal, Sierra Leone, Sudan, Tanzania, Tunisia, Uganda

[4] In version 4 of the SWIID one could download a spreadsheet with only the mean of the imputations. Using this summary dataset will obviously reintroduce the error. This option seems to have disappeared in SWIID 5.0.

And of course bad data may be multiplied by the imputation process or even amplified by an incorrect model. Even outside South Africa it is unlikely that measurement error will be entirely absent from some of the early inequality estimates. To that extent people relying on the 95% confidence intervals to capture all of the "associated uncertainty" with the Gini estimates will be building on quicksand. The uncertainty will be correctly measured only for countries and years where the original data and imputation model are reasonable.

References

Devereux, Stephen (1983) "South African Income Distribution, 1900-1980", ***SALDRU Working Paper*** 51, Cape Town: Southern African Labour and Development Research Unit, University of Cape Town. Available from www.opensaldru.uct.ac.za.

Jain, Shail (1975) ***Size Distribution of Income: A Compilation of Data***, Washington, D.C.: World Bank.

Jenkins, Stephen (forthcoming), "World Income Inequality Databases: an assessment of WIID and SWIID", ***Journal of Economic Inequality***

Lachman, Desmond and Kenneth Bercuson (eds), (1992), "Economic Policies for a New South Africa", ***IMF Occasional Paper*** 91, Washington: International Monetary Fund.

Lecaillon, Jacques, Felix Paukert, Christian Morrisson, and Dimitxi Germidis, (1984) ***Income Distribution and Economic Development: An Analytical Survey***, Geneva: International Labour Office.

McGrath, Michael (1984), "Inequality in the Size Distribution of Incomes in South Africa", ***DSS Staff Paper*** 2, Durban: Development Studies Unit, University of Natal. Available from http://opendocs.ids.ac.uk/opendocs/

Paukert, Felix (1973), "Income Distribution at Different Levels of Development: A Survey of Evidence", ***International Labour Review***, 108:97-125.

Simkins, Charles (1979), "The Distribution of Personal Income among Income Recipients, 1970 and 1976", ***DSRG Working Paper*** 9, Pietermaritzburg: Development Studies Research Group, University of Natal. Available from http://opendocs.ids.ac.uk/opendocs/

Solt, Frederick (2014a), "The Standardized World Income Inequality Database", Working Paper, SWIID Version 5.0, October 2014.

-------- (2014b), SWIID Version 5.0 web application, available at http://myweb.uiowa.edu/fsolt/swiid/swiid.html

-------- (2014c), "Using the Standardized World Income Inequality Database", documentation released with SWIID Version 5.0, October 2014.

UNU-WIDER (2014), "World Income Inequality Database (WIID 3.0A)", version 3A, June 2014. Available at http://www.wider.unu.edu/research/WIID-3a/en_GB/database/

Van der Berg, Servaas (1989), "Long Term Economic Trends and Development Prospects in South Africa", ***African Affairs***, 88:187-203.

About DataFirst

DataFirst is a data service dedicated to making South African and other African survey and administrative microdata available to researchers and policy analysts.

We promote high quality research by providing the essential research infrastructure for discovering and accessing data and by developing skills among prospective users, particularly in South Africa.

We undertake research on the quality and usability of national data and encourage data usage and data sharing.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708
info@data1st.org / support@data1st.org