Using the SWIID in Stata

Frederick Solt
Associate Professor of Political Science
University of Iowa
frederick-solt@uiowa.edu

The Standardized World Income Inequality Database (SWIID) takes a Bayesian approach to standardizing observations collected from the OECD Income Distribution Database, the Socio-Economic Database for Latin America and the Caribbean generated by CEDLAS and the World Bank, Eurostat, the World Bank's PovcalNet, the UN Economic Commission for Latin America and the Caribbean, national statistical offices around the world, and many other sources. Luxembourg Income Study data serves as the standard.

As described in Solt (2019), the SWIID maximizes the comparability of available income inequality data for the broadest possible sample of countries and years. But incomparability remains, and it is sometimes substantial. This remaining incomparability is reflected in the standard errors of the SWIID estimates, making it often crucial to take this uncertainty into account when making comparisons across countries or over time (Solt 2009, 238; Solt 2016, 14; Solt 2019). It was once the case that incorporating the standard errors into an analysis required considerable effort. It is now straightforward.

In version 8.0 of the SWIID, the inequality estimates and their associated uncertainty are represented by 100 draws from the posterior distribution: for any given observation, the differences across these imputations capture the uncertainty in the estimate. The `swiid8_0.zip` includes the file `swiid8_0.dta`, which is pre-formatted to facilitate taking this uncertainty into account. The following sections describe how to subset the data, merge in additional variables, and do analyses.

1 Getting Started

The `swiid8_0.dta` file is pre-formatted for use with Stata's tools for analyzing multiply imputed data. Estimates of each of four inequality measures and their associated uncertainty are represented by a placeholder variable (which has the measure's name but only missing data for all observations) plus 100 separate variables (prefixed with `_1_`, `_2_`, etc.): for any given observation, the differences across these 100 variables capture the uncertainty in the estimate.

The four measures are:

- `gini_disp`: Estimate of Gini index of inequality in equivalized (square root scale) household disposable (post-tax, post-transfer) income, using Luxembourg Income Study data as the standard.
- `gini_mkt`: Estimate of Gini index of inequality in equivalized (square root scale) household market (pre-tax, pre-transfer) income, using Luxembourg Income Study data as the standard.

- `abs_red`: Estimated absolute redistribution, the number of Gini-index points market-income inequality is reduced due to taxes and transfers: the difference between the `gini_mkt` and `gini_disp`.
- `rel_red`: Estimated relative redistribution, the percentage reduction in market-income inequality due to taxes and transfers: the difference between the `gini_mkt` and `gini_disp`, divided by `gini_mkt`, multiplied by 100.

```
use swiid8_0.dta, clear
```

2 Adding Variables

Generating new variables from the SWIID estimates requires a bit of care. To preserve Stata's recognition of how the SWIID is formatted for analysis, the `mi passive:` prefix must be used. Suppose we wanted to generate a variable for the log of `gini_net`. For this new variable to take into account the uncertainty in the SWIID estimates, instead of simply typing `gen ln_gini_net = ln(gini_net)`, we need to preface that command with the `mi passive:` prefix, as below:

```
mi passive: gen ln_gini_disp = ln(gini_disp)
```

The result is a placeholder variable for the new measure `ln(gini_net)`, plus 100 separate variables prefixed with `_1_`, `_2_`, etc. that together represent the uncertainty in our new measure. Note that there is no need to use `mi passive:` to create variables in the dataset that are not based on the SWIID estimates.

3 Merging

To merge the SWIID and additional data, simply merge the other dataset *into* the SWIID dataset. Note that this means that the SWIID should be the 'master' file in the merge, the other data should be the 'using' file.

Suppose we wanted to do a (simplified) replication of Solt, Habel, and Grant's (2011) analysis of World Values Survey data on religiosity. As our measure of religiosity, we will use the WVS item on respondents' self-report of the importance of God to their lives, which is measured on a ten-point scale. Given secularization theory, we will need to control for GDP per capita, which we will calculate from information from the Penn World Tables (Feenstra, Inklaar, and Timmer 2015). Below we first load the PWT dataset and use it to generate a dataset of GDP per capita (in thousands of dollars). Then we load the WVS data, generate our variables of interest, and merge in our PWT data. Finally, we merge these data into the SWIID.

```
// Get GDP per capita data from the Penn World Tables, Version 9.0 (Feenstra et al. 2015)
// download from http://www.rug.nl/research/ggdc/data/pwt/v90/pwt90.xlsx
// create gdppc and save as .dta

import excel using pwt90.xlsx, sheet("Data") firstrow clear
gen gdppc = rgdpe/pop/1000
drop if gdppc==.
```

```
keep country year gdppc
save pwt90_gdppc.dta, replace


// Get World Values Survey 6-wave data
// from http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp
// generate variables of interest, merge in the PWT data, and save

use WVS_Longitudinal_1981_2014_stata_v2015_04_18.dta, clear
kountry S003, from(iso3n)
rename NAMES_STD country
gen year = S020
gen country_year = S025
gen religiosity = F063 if F063>0
gen age = X003 if X003>0
gen educ = X025 if X025>0
gen male = (X001 == 1) if X001>0
keep country year country_year religiosity male educ age
merge m:1 country year using pwt90_gdppc.dta
drop if _merge!=3
drop _merge
save wvs_pwt.dta, replace

// Now merge these data *into* the SWIID
use swiid8_0.dta, clear

merge 1:m country year using wvs_pwt.dta
drop if _merge!=3
drop _merge
```

4 Analyzing

Once any additional variables are created or merged in, we may proceed to analysis. Continuing with our example, we estimate a three-level linear mixed-effects model of individual responses nested in country-years nested in countries using `mixed`. To take the uncertainty in the SWIID estimates into account, we construct our model comman as usual, but precede it with the `mi estimate:` prefix to perform it on each of the 100 variables that report the uncertainty in the SWIID estimates. Note that performing an analysis 100 times can be time-consuming.

```
mi estimate: mixed religiosity gini_disp gdppc age educ male || country: || country_year:
```

5 Working with Commands Unsupported by `mi estimate`

Unfortunately, the `mi estimate` does not support all estimation commands.[1] However, users can employ the `cmdok` option to work around this problem. Here's an example: We can use the `gmm` command (for general method of moments) even though it is not yet supported by the `mi estimate` prefix. Instead of specifying the model right after `mi estimate:`, we need to specify it after the `mi estimate, cmdok:` as following.

```
mi estimate, cmdok:gmm (religiosity - {b1}*gini_net - {b2}*educ - {b0}),
instruments(gini_net educ)
```

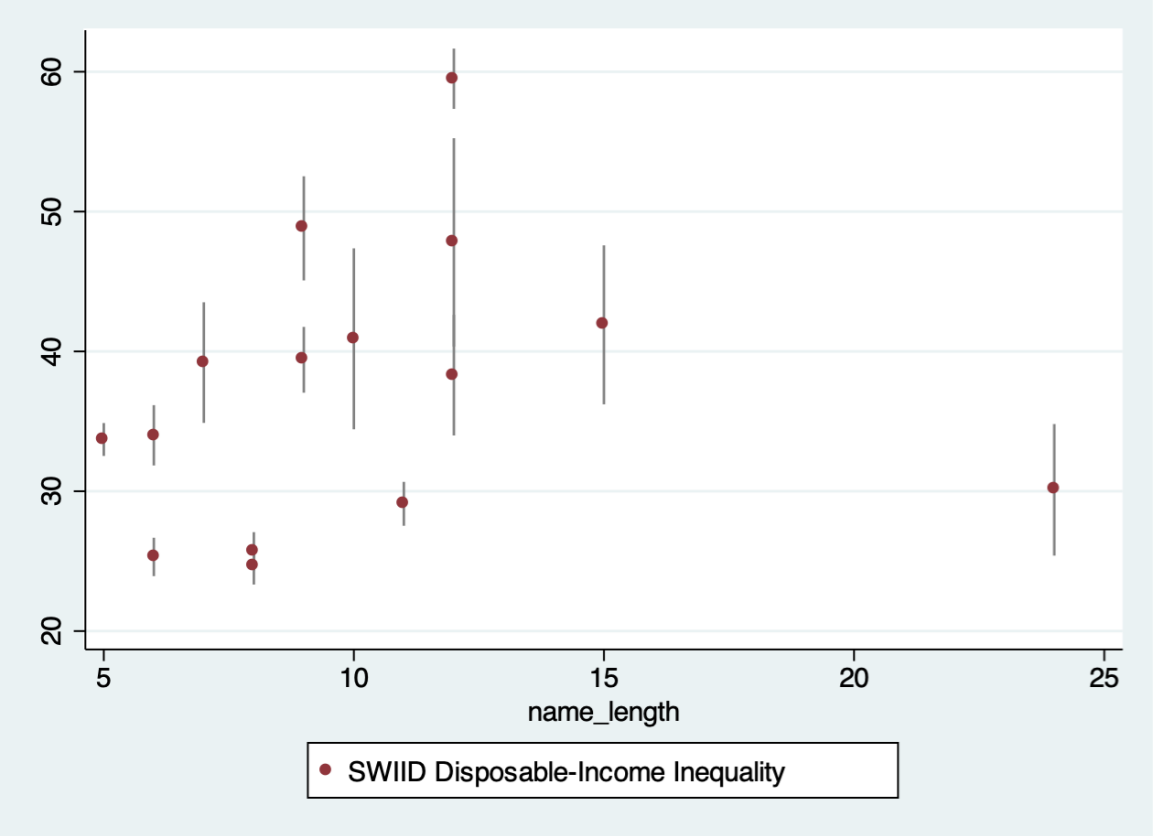
6 Mean-plus-standard-error Summary Format

The format described above facilitates taking the uncertainty in the SWIID estimates into account when conducting analyses. It does not, however, lend itself easily to tasks such plotting. The mean-plus-standard-error summary format is much better suited to such purposes. The SWIID is presented in this format in the `swiid7_1_summary.csv` file.

```
import delimited "swiid8_0_summary.csv", clear

// Calculate the bounds of the 95% uncertainty intervals
gen gini_disp_95ub = gini_disp + 1.96*gini_disp_se
gen gini_disp_95lb = gini_disp - 1.96*gini_disp_se

// A silly example
gen name_length = length(country)
gen first_letter = substr(country, 1, 1)
keep if year==2010 & first_letter=="S" /*2010 for Senegal, Serbia, . . .*/

// A scatterplot with 95% uncertainty intervals
twoway rspike gini_disp_95ub gini_disp_95lb name_length, lstyle(ci) || ///
    scatter gini_disp name_length, msize(small) ///
    legend(order(2 "SWIID Disposable-Income Inequality"))
```

[1] See the full list the prefix supports in the STATA document, mi estimation.

Figure 1: A Scatterplot with Confidence Intervals



7 Citing the SWIID

Please cite to the SWIID by referring to its article of record and including the version number and date of release:

Solt, Frederick. 2019. "Measuring Income Inequality Across Countries and Over Time: The Standardized World Income Inequality Database." SWIID Version 8.0, February 2019.

References

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2015. "The Next Generation of the Penn World Table." *American Economic Review* 105(10):3150–3182.

Solt, Frederick. 2009. "Standardizing the World Income Inequality Database." *Social Science Quarterly* 90(2):231–242.

Solt, Frederick. 2016. "The Standardized World Income Inequality Database." *Social Science Quarterly* 97(5):1267–1281. SWIID Version 8.0, February 2019.

Solt, Frederick. 2019. "Measuring Income Inequality Across Countries and Over Time: The Standardized World Income Inequality Database." SWIID Version 8.0, February 2019.

Solt, Frederick, Philip Habel, and J. Tobin Grant. 2011. "Economic Inequality, Relative Power, and Religiosity." *Social Science Quarterly* 92(2):447–465.