# 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) uses a custom missing-data multiple-imputation algorithm to standardize observations collected from the United Nations University's World Income Inequality Database version 2.0c, the OECD Income Distribution Database, the Socio-Economic Database for Latin America and the Caribbean generated by CEDLAS and the World Bank, the 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 (2016), 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 absolutely crucial to take this uncertainty into account when making comparisons across countries or over time (Solt 2009, p.238; Solt 2016, p.14). Using older versions of the SWIID, however, incorporating the standard errors into an analysis required considerable effort. It is now straightforward.

Beginning with 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. The `SWIIDv5_1.zip` includes the file `SWIIDv5_1.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 `SWIIDv5_1.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.

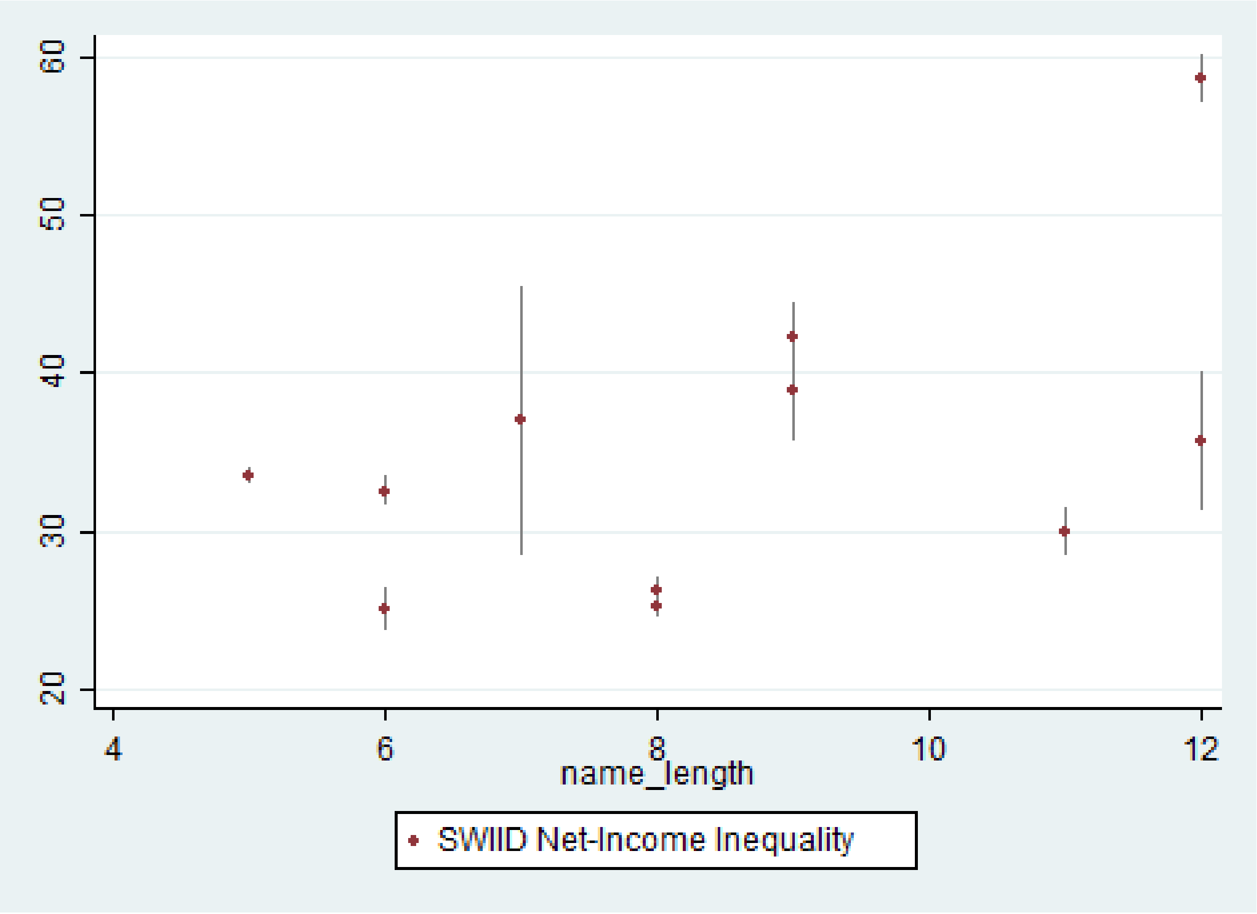
The four measures are:

- `gini_net`: 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_market`: 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_market` and `gini_net`.
- `rel_red`: Estimated relative redistribution, the percentage reduction in market-income inequality due to taxes and transfers: the difference between the `gini_market` and `gini_net`, divided by `gini_market`, multiplied by 100.

This format facilitates taking the uncertainty in its estimates into account when conducting analyses, as will be discussed below. 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 following code demonstrates how to put the SWIID in this summary format, as well as how to make a scatterplot with confidence intervals.

```
. use SWIIDv5_1.dta, clear
(SWIID v5.1, July 2016. Refer to the stata_swiid.pdf file for usage instruction
> s.)

. // Summarize the dataset
. keep country year _*

.
. foreach v in gini_net gini_market rel_red abs_red {
  2.     egen ‘v’ = rowmean(_*‘v’)
  3.     egen ‘v’_se = rowsd(_*‘v’)
  4.     gen ‘v’_95ub = ‘v’ + 1.96*‘v’_se
  5.     gen ‘v’_95lb = ‘v’ - 1.96*‘v’_se
  6. }
(2064 missing values generated)
(2064 missing values generated)
(2,064 missing values generated)
(2,064 missing values generated)
(2064 missing values generated)
(2064 missing values generated)
(2,064 missing values generated)
(2,064 missing values generated)

. drop _*

. sort country year

.
. // 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, . . .*/
(4,071 observations deleted)

.
. // A scatterplot with 95% confidence intervals
. twoway rspike gini_net_95ub gini_net_95lb name_length, lstyle(ci) || ///
>     scatter gini_net name_length, msize(small) ///
>     legend(order(2 "SWIID Net-Income Inequality"))

.
. graph export stata_scatter.png, replace //export the plot into a png file
(file stata_scatter.png written in PNG format)
```



## 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_net = ln(gini_net)
```

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 replicate Solt, Habel, and Grant's analysis (2011) of World Values Survey data on religiosity.

It requires variables from three datasets: 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).

To combine them into a dataset that STATA can analyze, we first need to 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/ggdc/docs/pwt90.dta
.
. use pwt90.dta, clear

. gen gdppc = rgdpe/pop/1000
(2,391 missing values generated)

. drop if gdppc==.
(2,391 observations deleted)

. keep country year gdppc
. save pwt90_gdppc.dta, replace
file pwt90_gdppc.dta saved

.
. // 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)

----------------------------------------
The command has finished.
The new variable is named NAMES_STD.
----------------------------------------

. rename NAMES_STD country

. gen year = S020

. gen religiosity = F063 if F063>0
(22,287 missing values generated)

. gen educ = X025 if X025>0
(45,129 missing values generated)

. keep country year religiosity educ

.
. merge m:1 country year using pwt90_gdppc.dta
(note: variable country was str22, now str34 to accommodate using data's
       values)

    Result                           # of obs.
    -----------------------------------------
    not matched                        62,617
        from master                    53,376  (_merge==1)
        from using                      9,241  (_merge==2)

    matched                           287,895  (_merge==3)
    -----------------------------------------

. drop if _merge!=3
(62,617 observations deleted)
```

```
. drop _merge

. save wvs_pwt.dta, replace
file wvs_pwt.dta saved

.
. // Now merge these data *into* the SWIID
. use SWIIDv5_1.dta, clear
(SWIID v5.1, July 2016. Refer to the stata_swiid.pdf file for usage instruction
> s.)

. merge 1:m country year using wvs_pwt.dta
(note: variable country was str30, now str34 to accommodate using data's
       values)
(note: variable year was int, now float to accommodate using data's values)

    Result                           # of obs.
    -----------------------------------------
    not matched                        48,849
        from master                     3,914  (_merge==1)
        from using                     44,935  (_merge==2)

    matched                           242,960  (_merge==3)
    -----------------------------------------

. drop if _merge!=3
(48,849 observations deleted)

. drop _merge
```

# 4 Analyzing

Once any additional variables are created or merged in, we may proceed to analysis. The original paper includes a three-level linear mixed-effect model of individual responses nested in country-years nested in countries. For illustration, we specify an over-simplified version of model at the individual level including only three variables: `religiosity` as the DV, `gini_net` as IV, and `educ` as control. To further reduce the computing time, we use the data only from 2014.

```
>  ...
NAMES_STD

. keep if year == 2014
(240,272 observations deleted)
```

```
. mi xtset, clear // clean the board
```

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.

```
>  ...
NAMES_STD

. mi estimate: reg religiosity gini_net educ

Multiple-imputation estimates                   Imputations       =        100
Linear regression                               Number of obs     =      2,666
                                                Average RVI       =     0.3083
                                                Largest FMI       =     0.1354
                                                Complete DF       =       2663
DF adjustment:   Small sample                   DF:     min       =   1,617.27
                                                        avg       =   2,001.62
                                                        max       =   2,661.00
Model F test:       Equal FMI                   F(   2, 2405.9)   =      24.62
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------
 religiosity |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gini_net |    .036713   .0101112     3.63   0.000     .0168806    .0565455
        educ |  -.0531121   .0133782    -3.97   0.000    -.0793449   -.0268794
       _cons |   8.064514   .4652991    17.33   0.000     7.151904    8.977123
------------------------------------------------------------------------------
```

## 5 Working with Unsupported Commands by `mi estimate`

Unfortunately, the `mi estimate` does not support all estimation commands.[1] However, users can apply the `cmdok` option to get around such problem. Here's an example: Let's replicate the preceding example with the `gmm` command (for general method of moments), a build-in command in STATA yet unsupported 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.

```
>  ...
NAMES_STD
```

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

```
. mi estimate, cmdok:gmm (religiosity - {b1}*gini_net - {b2}*educ - {b0}), inst
> ruments(gini_net educ)

Multiple-imputation estimates                   Imputations       =        100
                                                Number of obs     =      2,666
                                                Average RVI       =     0.3354
                                                Largest FMI       =     0.1200
DF adjustment:   Large sample                   DF:     min       =   6,899.68
                                                        avg       =   4.50e+58
Within VCE type:       Robust                           max       =   1.35e+59

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         /b1 |    .036713   .0107316     3.42   0.001     .0156759    .0577502
         /b2 |  -.0531121   .0135038    -3.93   0.000    -.0795791   -.0266452
         /b0 |   8.064514   .4961315    16.25   0.000     7.091976    9.037051
------------------------------------------------------------------------------
```

As shown above, we successfully replicated the results in the OLS estimation with an `mi estimate` unsupported command.

There is a noteworthy point for applying this method, though. Be aware that, for applying the Rubin's combination rules, STATA needs at least four parts of information: the global macro `e(cmd)`, the coefficients matrix `e(b)`, the variance–covariance matrix matrix `e(V)`, and the residual degrees of freedom `e(df_r)`. Therefore, one cannot apply the above methods when the output of a command does not store these matrices. See more techical discussion about this situation in the STATA FAQ document for `cmdok`.

## 6 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. 2016. "The Standardized World Income Inequality Database." *Social Science Quarterly* 97. SWIID Version 5.1, July 2016.

## References

Feenstra, Robert C, Robert Inklaar, and Marcel P Timmer. 2015. "The Next Generation of the Penn World Table." *The 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.

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