sgini

Generalized Gini, Concentration coefficients, Factor decomposition, and Gini correlations in Stata

Philippe Van Kerm
University of Luxembourg and Luxembourg Institute for Socio-Economic Research[‡]

September 2009 (last revised April 2020)

Abstract `sgini` is a light-weight user-written Stata package to compute generalized Gini and concentration coefficients, along with their factor decomposition (or decomposition by source). A companion command provides calculations of Gini correlations. This manual describes syntax, formulas and usage examples.

Keywords `sgini` ; Stata ; generalized Gini ; Concentration coefficient ; Gini correlation ; factor decomposition

JEL Classification C88; D31

1 Introduction

`sgini` is a user-written Stata command for calculation of generalized Gini—also known as S-Gini—and Concentration coefficients from unit-record data. `sgini` computes classic relative (scale invariant) Gini indices of inequality by default but can be requested to produce absolute (translation invariant) indices or aggregate welfare S-Gini indices. The command can also calculate decomposition of these indices by factor components (income sources). As this document shows, while the scope of the command itself is seemingly relatively narrow, `sgini` can serve as a flexible building block in a wide array of applications. As a by-product, a companion command provides calculations of Gini correlations.

The command is available online for installation in net-aware Stata.[1] At the command prompt, type

```
ssc install sgini
```

2 Covariance-based expressions for generalized Gini, Concentration coefficients and Gini correlations

Gini and concentration coefficients

The Gini coefficient is the most popular measure of inequality. One of its many formulations (see Yitzhaki, 1998) is based on a covariance expression:

$$\text{GINI}(X) = -2\ \text{Cov}\left(\frac{X}{\mu(X)},\ (1-F(X))\right)$$

where X is a random variable of interest with mean $\mu(X)$, and $F(X)$ is its cumulative distribution function (see, e.g., Lerman & Yitzhaki, 1984, Jenkins, 1988).[2]

[‡]Maison des Sciences Humaines, Campus Belval, 11 Porte des Sciences, L-4364 Esch/Alzette, Luxembourg. E-mail: philippe.vankerm@liser.lu.

[1]The latest version of the `sgini` package is 2.0.0 (of 2020-04-23); Stata 8.2 or later is required (Van Kerm, 2020).

[2]The Gini coefficient is perhaps more generally known as one minus twice the area under the Lorenz curve of X which, if X is income, plots the share of total income held by the poorest $100 \times p$ percent of the population against p.

Closely related to the Gini coefficient is the Concentration coefficient. The Concentration coefficient measures the association between two random variables and can be expressed as

$$\text{CONC}(X, Y) = -2 \ \text{Cov}\left(\frac{X}{\mu(X)}, \ (1 - G(Y))\right)$$

where $G(Y)$ is the cumulative distribution function of Y. $\text{CONC}(X, Y)$ reflects how much X is concentrated on observations with high ranks in Y (see, e.g., Kakwani, 1977a).

Single-parameter generalization

A single-parameter generalization of the Gini coefficient has been proposed by Donaldson & Weymark (1980, 1983) and Yitzhaki (1983). The generalized Gini coefficient (a.k.a. the S-Gini, or extended Gini coefficient) can also be expressed as a covariance:

$$\text{GINI}(X; \upsilon) = -\upsilon \ \text{Cov}\left(\frac{X}{\mu(X)}, \ (1 - F(X))^{\upsilon - 1}\right)$$

where υ is a parameter tuning the degree of 'aversion to inequality'. The standard Gini corresponds to $\upsilon = 2$. See Yitzhaki & Schechtman (2005) for a recent review.

The generalized Concentration coefficient can be similarly defined as

$$\text{CONC}(X, Y; \upsilon) = -\upsilon \ \text{Cov}\left(\frac{X}{\mu(X)}, \ (1 - G(Y))^{\upsilon - 1}\right).$$

Aggregate welfare and absolute Gini coefficients

While measures of inequality are typically taken as relative (or scale invariant)—one considers the distribution of, say, income shares, so that inequality is unaffected by an equi-proportionate change in all data—there are situations in which there is either interest in taking levels into account too, a.k.a. aggregate 'welfare' indices (Sen, 1976), or in considering deviations from the mean, a.k.a. 'absolute' inequality (Blackorby & Donaldson, 1980). In the first case, Concentration coefficients (and Gini coefficients with $X = Y$) are simply redefined as

$$\begin{aligned} \text{AGGCONC}(X, Y; \upsilon) &= \mu(X) \ (1 - \text{CONC}(X, Y; \upsilon)) \\ &= \mu(X) + \upsilon \ \text{Cov}\left(X, \ (1 - G(Y))^{\upsilon - 1}\right). \end{aligned}$$

$\text{AGGCONC}(X, Y; \upsilon)$ can be interpreted as the "equally distributed equivalent" mean of X. The second measure is obtained as

$$\begin{aligned} \text{ABSCONC}(X, Y; \upsilon) &= \mu(X) - \text{AGGCONC}(X, Y; \upsilon) \\ &= \mu(X) \ \text{CONC}(X, Y; \upsilon). \end{aligned}$$

$\text{ABSCONC}(X, Y; \upsilon)$ is a measure of inequality that is invariant to equal additions to all data.

Gini correlation

Concentration coefficients capture the association between two random variables, and this naturally leads to a measure of correlation known as the 'Gini correlation' (Schechtman & Yitzhaki, 1987, 1999). The class of generalized Gini correlation coefficients is defined as

$$\begin{aligned} R(X, Y; \upsilon) &= \frac{\text{Cov}\left(X, \ (1 - G(Y))^{\upsilon - 1}\right)}{\text{Cov}\left(X, \ (1 - F(X))^{\upsilon - 1}\right)} \\ &= \frac{\text{CONC}(X, Y; \upsilon)}{\text{GINI}(X; \upsilon)}. \end{aligned}$$

Its properties can be understood as a mixture of Pearson's and Spearman's correlations. Note that unlike these two correlation measures, the Gini correlation coefficient is not symmetric: $R(X, Y; \upsilon) \neq R(Y, X; \upsilon)$. See Schechtman & Yitzhaki (2003) for a discussion of the properties of generalized Gini correlation coefficients and their applications.

3 Calculation

Covariance-based expressions for the generalized Gini and Concentration coefficients are convenient for calculations from unit-record data. Estimation simply involves estimating a sample covariance between the observations from variable X (divided by their sample mean) and the (fractional) ranks of observations from variable X or Y.

The only delicate step is the computation of fractional ranks in the presence of tied data and/or sampling weights. This can be critical when dealing with ordinal data. Consider a sample of N observations on a variable Y with associated sampling weights: $\{(y_i, w_i)\}_{i=1}^{N}$. Let K be the number of distinct values observed on Y, denoted $y_1^* < y_2^* < \ldots < y_K^*$, and denote by π_k^* the corresponding weighted sample proportions with any value y_k^*:

$$\pi_k^* = \frac{\sum_{i=1}^{N} w_i \mathbf{1}(y_i = y_k^*)}{\sum_{i=1}^{N} w_i}$$

($\mathbf{1}(condition)$ is equal to 1 if *condition* is true and 0 otherwise). The fractional rank attached to each y_k^* is then given by

$$F_k^* = \sum_{j=0}^{k-1} \pi_j^* + 0.5\pi_{j+1}^*$$

where $\pi_0^* = 0$ (Lerman & Yitzhaki, 1989, Chotikapanich & Griffiths, 2001). Each observation in the sample is then associated with the fractional rank

$$F_i = \sum_{k=1}^{K} F_k^* \mathbf{1}(y_i = y_k^*).$$

This procedure ensures that tied observations are associated with identical fractional ranks and that the sample mean of the fractional ranks is equal to 0.5. $\{(F_i, y_i, w_i)\}_{i=1}^{N}$ can then be plugged in a standard sample covariance formula. This makes the resulting Gini or concentration coefficient estimate independent on sample size and ordering of the data.[3]

4 Applications

Gini coefficients are popular measures of inequality. Similarly, Concentration coefficients are often used to measure income-related inequalities in other socially important variables. van Doorslaer *et al.* (1997), for example, compared income-related inequalities in health across a number of countries using the Concentration coefficient of a self-reported health measure against income.

Both measures are also used as building blocks for a number of related applications. Four cases are illustrated here: (i) the decomposition of income inequality by sources, (ii) the measurement of tax progressivity and horizontal equity, (iii) the measurement of income mobility and of the 'pro-poorness' of growth, and (iv) the measurement of income polarization.

[3] See Yitzhaki & Schechtman (2005), Berger (2008), Chen & Roy (2009), and Davidson (2009) for recent discussions of this issue. Also see Cox (2002) for a more general discussion of (fractional) ranks. In Stata, a common mistake is to use `cumul` or `glcurve`'s `pvar()` option to compute fractional ranks. The former returns empirical CDF estimates at sample points and the latter saves plotting coordinates for generalized Lorenz curves. None of these is adequate for plugging in covariance-based expressions of concentration coefficients.

4.1 Decomposition of income inequality by source

Total family income can be seen as the sum of a number of components: earnings, capital income, transfer income, etc. There might be interest in identifying the contribution of each of these sources to inequality in total income. The 'natural' decomposition of Generalized Gini coefficients is

$$\text{GINI}(Y;\upsilon) = \sum_{k=1}^{K} \frac{\mu(Y^k)}{\mu(Y)} \times \text{CONC}(Y^k, Y;\upsilon)$$

where $\text{CONC}(Y^k, Y;\upsilon)$ is the generalized Concentration coefficient of incomes from source k with respect to total income and $\mu(Y^k)$ and $\mu(Y)$ denote means of source k and total income respectively (Fei *et al.*, 1978, Lerman & Yitzhaki, 1985). Lerman & Yitzhaki (1985) also noted that $\text{CONC}(Y^k, Y;\upsilon)$ can further be expressed as

$$\text{CONC}(Y^k, Y;\upsilon) = \text{GINI}(Y^k;\upsilon) \times \text{R}(Y^k, Y;\upsilon)$$

where $\text{R}(Y^k, Y;\upsilon)$ is the '(generalized) Gini correlation'

$$\text{R}(Y^k, Y;\upsilon) = \frac{\text{Cov}\left(Y^k, (1-F(Y))^{\upsilon-1}\right)}{\text{Cov}\left(Y^k, (1-F^k(Y^k))^{\upsilon-1}\right)}.$$

Finally, Lerman & Yitzhaki (1985) derive an expression for the relative impact on the Gini coefficient of a marginal increase in the size of source k:

$$\begin{aligned}\frac{1}{\text{GINI}(Y;\upsilon)} \times \frac{\partial \text{GINI}(Y;\upsilon)}{\partial e^k} &= \frac{\mu(Y^k)}{\mu(Y)}\left(\frac{\text{CONC}(Y^k, Y;\upsilon)}{\text{GINI}(Y;\upsilon)} - 1\right) \\ &= \frac{\mu(Y^k)}{\mu(Y)}\left(\frac{\text{GINI}(Y^k;\upsilon)\text{R}(Y^k, Y;\upsilon)}{\text{GINI}(Y;\upsilon)} - 1\right)\end{aligned}$$

Note how these components are combinations of estimates of Gini and Concentration coefficients, along with sample means.

López-Feldman (2006) discusses this decomposition in greater details and describes `descogini`, a Stata command for calculating the components of the decomposition.[4]

4.2 Tax progressivity and horizontal equity

Much of the analyses on taxation schemes attempt to measure how 'progressive' is a tax schedule, that is, how much inequality is reduced after application of the tax. A popular measure is the Reynolds-Smolensky index of redistributive effect defined as the difference between the Gini coefficient of pre-tax income and the Gini coefficient of post-tax income (Reynolds & Smolensky, 1977):

$$\Pi^{RS} = \text{GINI}(X^{\text{pre}}) - \text{GINI}(X^{\text{post}})$$

where X^{pre} and X^{post} are pre- and post-tax income, respectively. The Kakwani measure of progressivity is similarly defined (Kakwani, 1977b):

$$\Pi^{K} = \text{CONC}(T, X^{\text{pre}}) - \text{GINI}(X^{\text{pre}})$$

where T is the tax paid: $T = X^{\text{pre}} - X^{\text{post}}$. Combining the progressivity measure with a component capturing the re-ranking induced by the tax schedule leads to a decomposition of Π^{RS} as

$$\Pi^{RS} = \frac{g}{1-g}\Pi^{K} - R$$

[4] `descogini` does not currently handle sample weights (as of the January 2008 version) and is limited to $\upsilon = 2$. `sgini` can optionally be used to report similar decomposition coefficients without these constraints (see *supra*).

where $\mathsf{R} = (\mathsf{CONC}(X^{\text{post}}, X^{\text{pre}}) - \mathsf{GINI}(X^{\text{post}}))$ captures the effect of re-ranking on the net reduction in the Gini coefficient, and g is the average tax rate. See Lambert (2001) for a textbook exposition. Again, all that is required to compute these measures is estimation of a Gini and Concentration coefficients as calculated by `sgini`.

These (and other) measures are also computed by the Stata command `progres` available on SSC (Peichl & Van Kerm, 2007).

4.3 Income mobility and pro-poor growth

Transposing concepts from the progressivity measurement, Jenkins & Van Kerm (2006) relate the change in income inequality over time to the progressivity of individual income growth—a measure of the 'pro-poorness' of economic growth—and mobility in the form of re-ranking:

$$\Delta(\upsilon) = \mathsf{R}(\upsilon) - \mathsf{P}(\upsilon)$$

where

$$\mathsf{P}(\upsilon) = \text{GINI}(X^0; \upsilon) - \text{CONC}(X^1, X^0\upsilon)$$

and

$$\mathsf{R}(\upsilon) = \text{GINI}(X^1; \upsilon) - \text{CONC}(X^1, X^0; \upsilon)$$

$\mathsf{P}(\upsilon)$ can be interpreted as an indicator of how much growth has benefited disproportionately to individuals at the bottom of the distribution in the initial time period. $\mathsf{R}(\upsilon)$ captures how much a progressive income growth has lead to re-ranking between individuals, so that the net reduction in inequality is the difference between $\mathsf{P}(\upsilon)$ and $\mathsf{R}(\upsilon)$. Note that $\mathsf{R}(\upsilon)$ can also be interpreted as a measure of mobility (in the form of re-ranking) in its own right (Yitzhaki & Wodon, 2004). In an analysis of cross-country convergence in GDP, O'Neill & Van Kerm (2008) have interpreted $\Delta(\upsilon)$ as a measure of 'σ-convergence' and $\mathsf{P}(\upsilon)$ as a measure of 'β-convergence', thereby reconciling the two concepts in a single framework.

This decomposition of changes over time in the Gini coefficient is implemented in Stata in the `dsginideco` package available from the SSC archive (Jenkins & Van Kerm, 2009).

The literature on income mobility also uses Gini and concentration coefficients. Jenkins & Van Kerm (2016) proposes to assess individual income growth as

$$\text{M1}(\upsilon) = \text{TOTCONC}(\mathsf{Z}, X^0; \upsilon)$$

where Z is a measure of individual (or household-level) income change, the simplest of which is $\mathsf{Z} = (Y_1 - Y_0)$, or $\mathsf{Z} = (\ln(Y_1) - \ln(Y_0))$. Demuynck & van de Gaer (2012) advocates instead

$$\text{M2}(\upsilon) = \text{TOTGINI}(\mathsf{Z}; \upsilon).$$

$\text{M1}(\upsilon)$ and $\text{M2}(\upsilon)$ differ in how individuals are ranked when calculating (fractional) ranks: by initial income in M1 and by Z in M2.

4.4 Income polarization

The Gini coefficient has also been used as a building block of measures of income "bipolarization". Arrange a population in increasing order of income and divide it in two equal-sized groups: the 'poor' are individuals with an income below the median and the 'rich' are those with income above the median. Measures of bipolarization capture the distance between these two groups. Denote the median $\text{M}(Y)$, mean income $\mu(Y)$, mean income among the poor $\mu(Y^{\mathsf{P}})$, mean income among the rich $\mu(Y^{\mathsf{R}})$, the Gini coefficient among the poor $\text{GINI}(Y^{\mathsf{P}})$, the Gini coefficient among the rich $\text{GINI}(Y^{\mathsf{R}})$ and the overall Gini, $\text{GINI}(Y)$. Silber *et al.* (2007) show that several measures of bipolarization can

be expressed in terms of 'within-group Gini' and 'between-group Gini' which can be written in this situation as:

$$\text{Within}(Y^P, Y^R) = \frac{1}{4}\left(\frac{\mu(Y^P)}{\mu(Y)}\text{GINI}(Y^P) + \frac{\mu(Y^R)}{\mu(Y)}\text{GINI}(Y^R)\right)$$

and

$$\text{Between}(Y^P, Y^R) = \frac{1}{4}\left(\frac{\mu(Y^R)}{\mu(Y)} - \frac{\mu(Y^P)}{\mu(Y)}\right).$$

Note that the 'between-group Gini' is equivalent to estimating the Gini coefficient of mean income in the two groups, that is $\text{GINI}((\mu(Y^P), \mu(Y^R))')$. The bipolarization index suggested by Silber *et al.* (2007) is defined as

$$P_1 = \frac{\text{Between}(Y^P, Y^R) - \text{Within}(Y^P, Y^R)}{\text{GINI}(Y)},$$

the index of Wolfson (1994) as

$$P_2 = \left(\text{Between}(Y^P, Y^R) - \text{Within}(Y^P, Y^R)\right)\frac{\mu(Y)}{\text{Med}(Y)},$$

and the index proposed by Zhang & Kanbur (2001) as

$$P_3 = \frac{\text{Between}(Y^P, Y^R)}{\text{Within}(Y^P, Y^R)}.$$

See Silber *et al.* (2007). `sgini` makes estimation of these bipolarization measures straightforward (see examples below). It also opens possibilities for extensions of these measures by using variations of the inequality aversion parameter υ.[5]

5 The `sgini` command

`sgini` is a light-weight command to compute generalized Gini and concentration coefficents from unit-record data in Stata (based on the formulae from Section 2). `sgini` can also report a decomposition by source—the factor decomposition. `sgini` comes with two companion commands: `fracrank` for generating fractional ranks as described in Section 3 and `sginicorr` for calculation of generalized Gini *correlation* coefficients.

`sgini` does not provide sampling variance estimates (as an r-class command) but it is easily bootstrapped using a `bootstrap` or `svy bootstrap` prefix.

5.1 Syntax

`sgini` *varlist* [*if*] [*in*] [*weight*] [, `parameters(`*numlist*`)` `sortvar(`*varname*`)`
`fracrankvar(`*varname*`)` `sourcedecomposition` `absolute` `aggregate` `welfare` `format(`*%fmt*`)`
]

`fweight`, `aweight`, and `pweight` are allowed; see [U] **11.1.6 weight – Weights**.
`by`, `bootstrap`, `jackknife` are allowed; see [U] **11.1.10 Prefix commands**.
Time-series operators are accepted; see [U] **11.4.3 Time-series varlists**.

`sgini` computes the generalized Gini or concentration coefficients for each variable in *varlist* according to the parameters passed in option `parameters` and with ordering variable of option `sortvar`. Decomposition by income source is optionally computed. See detailed option descriptions below.

[5] The user-written command `bipolar`, available on the SSC archive, calculates these indicators (Fusco & Van Kerm, 2020).

Multiple variables and multiple inequality aversion parameters can be passed to `sgini`. Beware that if multiple variables are input, `sgini` discards observations with missing data on *any* of the input variables and computes all coefficients on the resulting sample.

`fracrank` *varname* [*if*] [*in*] [*weight*] `, generate(`*newvarname*`)`

`fweight`, `aweight`, and `pweight` are allowed; see [U] **11.1.6 weight − Weights**.
Time-series operators are accepted; see [U] **11.4.3 Time-series varlists**.

`fracrank` takes one numeric variable as input and creates a new variable filled with the corresponding fractional rank for each observation.

`sginicorr` *varlist* [*if*] [*in*] [*weight*] [`, parameter(`*real*`) format(`*%fmt*`)`]

`fweight`, `aweight`, and `pweight` are allowed; see [U] **11.1.6 weight − Weights**.
`by`, `bootstrap`, `jackknife` are allowed; see [U] **11.1.10 Prefix commands**.
Time-series operators are accepted; see [U] **11.4.3 Time-series varlists**.

`sginicorr` computes the generalized Gini correlation matrix between all variables in *varlist* with sensitivity parameter passed in `param` (default is 2). Case-wise deletion is applied; `sginicorr` discards observations with missing data on *any* of the variables in *varlist*.

5.2 Options

5.2.1 `sgini` options

`parameter(`*numlist*`)` specifies generalized Gini parameters. Default is 2, leading to the standard Gini or Concentration coefficient. Multiple parameters can be specified.

`sortvar(`*varname*`)` requests the computation of a Concentration coefficient by cumulating the variable(s) of interest in increasing order of *varname*. Default is to cumulate the variable(s) of interest against themselves, leading to Gini coefficients.

`aggregate` and `absolute` request, respectively, computation of aggregate S-Gini welfare measures or computation of absolute Gini and Concentration coefficients, instead of the relative inequality measures. They are mutually exclusive and incompatible with `sourcedecomposition`. `welfare` is synonymous for `aggregate`.

`fracrankvar(`*varname*`)` is a rarely used option that passes the name of an existing variable *varname* containing pre-specified fractional ranks based on which the Gini and Concentration coefficients can be computed. This is a potentially dangerous option, but it may lead to considerable speed gains under certain circumstances. It is essential that the fractional rank variable be computed correctly in the first place (using e.g., `fracrank`) and on the adequate sample (think missing data, `if` clauses, ordering).

`sourcedecomposition` requests factor decomposition of indices. It is relevant when more than one variable is passed in *varlist*. It requests that a variable be created by taking the row sum of all elements in *varlist* computes the Gini (or Concentration) coefficient for this new variable, and estimates the contribution of each element of *varlist* to the latter by applying the "natural" decomposition rule for Gini coefficients (as in Lerman & Yitzhaki, 1985).

`format(`*%fmt*`)` controls the display format; default is %4.3f.

5.2.2 `fracrank` options

`generate(`*newvarname*[`,replace`]`)` specifies the name for the created variable.

5.2.3 sginicorr options

parameter(*real*) specifies generalized Gini parameters. Default is 2 leading to the standard Gini correlation coefficient.

format(*%fmt*) controls the display format; default is %4.3f.

5.3 Saved results

sgini is an r-class command and saves the following results:

Scalars

r(N)	number of observations
r(sum_w)	sum of weights
r(coeff)	estimated coefficient for first variable, first parameter

Matrices

r(coeffs)	all estimated coefficients vector
r(parameters)	inequality aversion parameters vector
r(r)	Gini correlations between source and total income (if requested)
r(c)	concentration coefficients of each source (if requested)
r(elasticity)	elasticities between source and total Gini (if requested)
r(s)	factor shares (if requested)

Macros

r(varlist)	*varlist*
r(paramlist)	list of parameters from option param
r(sortvar)	*varname* if sortvar(*varname*) specified

sginicorr saves the following results:

Scalars

r(N)	number of observations
r(sum_w)	sum of weights
r(rho)	Gini correlation $\mathsf{R}(\mathsf{X},\mathsf{Y})$ between first (X) and second (Y) variable in *varlist*
r(param)	sensitivity parameter

Matrices

r(Rho)	Gini correlation matrix

Macros

r(varlist)	*varlist*

5.4 Dependencies on user-written packages

sgini does not require other user-written packages. sginicorr requires sgini.

6 Examples

The following examples illustrate usage of `sgini` using data from the National Longitudinal Survey of Youth, available from the Stata Press website. Take this as merely illustrative of syntax using data easily available from within Stata, not as of any substantive interest!

First open the data, generate a wage variable and `xtset` the data as appropriate.

```
. use http://www.stata-press.com/data/r9/nlswork , clear
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. xtset idcode year
       panel variable:  idcode (unbalanced)
        time variable:  year, 68 to 88, but with gaps
                delta:  1 unit
. gen w = exp(ln_wage)
```

`sgini` can then be used to estimate Gini coefficients on the wage variable (across all waves of data). The second syntax produces 'absolute' Gini coefficients for multiple inequality aversion parameters

```
. sgini w
Gini coefficient for w
```

Variable	v=2
w	0.2732

```
. sgini w , parameters(1.5(.5)4) absolute
Generalized Gini coefficient for w
```

Variable	v=1.5	v=2	v=2.5	v=3	v=3.5	v=4
w	1.1207	1.6522	1.9812	2.2113	2.3844	2.5213

Multiple variables can be passed in *varlist* and `sgini` accepts time-series variables, here for estimating concentration coefficients with the `sortvar(`*varname*`)` option. (Note how results for variable `w` differ from the first example because of the case-wise deletion of observations with missing data on `L.w` and `L2.w`.)

```
. sgini w L.w L2.w
Gini coefficient for w, L.w, L2.w
```

Variable	v=2
w	0.2023
L.w	0.1918
L2.w	0.1926

```
. sgini w L.w L2.w  , sortvar(w) param(2 3)
Generalized Concentration coefficient for w, L.w, L2.w against w
```

Variable	v=2	v=3
w	0.2023	0.2836
L.w	0.1581	0.2198
L2.w	0.1387	0.1921

```
. return list
scalars:
              r(sum_w) =  3481
                  r(N) =  3481
              r(coeff) =  .2023171625445915
macros:
            r(sortvar) : "w"
          r(paramlist) : "2 3"
            r(varlist) : "w L.w L2.w"
matrices:
             r(coeffs) :  1 x 6
         r(parameters) :  1 x 2
. matrix list r(coeffs)
r(coeffs)[1,6]
          param1:    param1:    param1:    param2:    param2:    param2:
                          L.         L2.                  L.         L2.
               w          w          w          w          w          w
Coeff  .20231716   .1581236  .13867147  .28359513  .21983376  .19212248
```

`sgini` is also 'by-able':

```
. sort year
. by year: sgini w
```

```
-> year = 68
Gini coefficient for w
```

Variable	v=2
w	0.1969

```
-> year = 69
Gini coefficient for w
```

Variable	v=2
w	0.1971

```
-> year = 70
```

(*output omitted*)

```
-> year = 88
Gini coefficient for w
```

Variable	v=2
w	0.3553

Specifying option `sourcedecomposition` produces the factor decomposition. Note that it is assumed that all variables in *varlist* is a source. They are aggregated to a total (row) variable. The following example therefore examines inequality in the observation-specific *sum* of wages for years 1986, 1987 and 1988 and decomposes the total into contributions of each year. (A more classic example would, for example, decompose total income into contributions of, say, earnings, rents and social

benefits.)

```
. sort idcode year

. sgini w L.w L2.w  if year==73 , sourcedecomposition

Gini coefficient for w, L.w, L2.w

------------------------
    Variable |       v=2
-------------+----------
           w |    0.2150
         L.w |    0.2077
        L2.w |    0.2043
------------------------

Decomposition by source:
  TOTAL =  w +  L.w +  L2.w

Parameter: v=2

------------------------------------------------------------------------------
    Variable |     Share   Coeff.    Corr.    Conc.  Contri. %Contri. Elasticity
             |         s         g         r     c=g*r    s*g*r  s*g*r/G  s*g*r/G-s
-------------+----------------------------------------------------------------
           w |    0.3492    0.2150    0.9427    0.2027   0.0708   0.3634    0.0142
         L.w |    0.3350    0.2077    0.9521    0.1978   0.0663   0.3402    0.0052
        L2.w |    0.3158    0.2043    0.8950    0.1828   0.0577   0.2964   -0.0194
-------------+----------------------------------------------------------------
       TOTAL |    1.0000    0.1948    1.0000    0.1948   0.1948   1.0000    0.0000
------------------------------------------------------------------------------

. return list

scalars:
              r(sum_w) =  1051
                  r(N) =  1051
              r(coeff) =  .194756281863436

macros:
           r(paramlist) : "2"
             r(varlist) : "w L.w L2.w"

matrices:
                  r(s) :  1 x 3
         r(elasticity) :  1 x 3
         r(relcontrib) :  1 x 3
            r(contrib) :  1 x 3
                  r(c) :  1 x 3
                  r(r) :  1 x 3
             r(coeffs) :  1 x 4
         r(parameters) :  1 x 1
```

A Gini correlation matrix can be computed by `sginicorr`.

```
. sginicorr w L.w L2.w
Gini correlation matrix (v=2):
                  L.     L2.
           w       w       w
   w  1.0000  0.7885  0.6704
 L.w  0.8245  1.0000  0.7716
L2.w  0.7201  0.7757  1.0000

(Note: Gini correlations are asymmetric. In corr(F(X),Y), X is the row variable and Y is t
> he column variable.)

. return list

scalars:
           r(parameter) =  2
                 r(rho) =  .7885315864149651
              r(sum_w) =  3481
                  r(N) =  3481
```

```
macros:
            r(varlist) : "w L.w L2.w"
matrices:
               r(Rho) :  3 x 3
```

Stata's built-in `bootstrap` or `jackknife` prefix commands can be used with `sgini` to produce standard errors and confidence intervals.

```
. bootstrap G=r(coeff) , reps(250) nodots : sgini w if !mi(w) & year==88
Warning:  Because sgini is not an estimation command or does not set e(sample),
          bootstrap has no way to determine which observations are used in calculating
          the statistics and so assumes that all observations are used.  This means that
          no observations will be excluded from the resampling because of missing values
          or other reasons.

          If the assumption is not true, press Break, save the data, and drop the
          observations that are to be excluded.  Be sure that the dataset in memory
          contains only the relevant data.
Bootstrap results                                 Number of obs     =      2,272
                                                  Replications      =        250

      command:  sgini w
            G:  r(coeff)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           G |   .3552972   .0095474    37.21   0.000     .3365847    .3740097
------------------------------------------------------------------------------

. jackknife G=r(coeff) , rclass nodots: sgini w if !mi(w) & year==88
Jackknife results                                 Number of obs     =      2,272
                                                  Replications      =      2,272

      command:  sgini w if !mi(w) & year==88
            G:  r(coeff)
         n():  r(N)

------------------------------------------------------------------------------
             |             Jackknife
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           G |   .3552972   .0100737    35.27   0.000     .3355426    .3750519
------------------------------------------------------------------------------
```

Note that (sampling) weights can be taken into account with `bootstrap`'s `force` option. (Assuming here that the variable `hours` was some sampling weight variable.)

```
. bootstrap G=r(coeff) , reps(250) nodots force : sgini w [aw=hours] if !mi(w) & year==88
Warning:  Because sgini is not an estimation command or does not set e(sample),
          bootstrap has no way to determine which observations are used in calculating
          the statistics and so assumes that all observations are used.  This means that
          no observations will be excluded from the resampling because of missing values
          or other reasons.

          If the assumption is not true, press Break, save the data, and drop the
          observations that are to be excluded.  Be sure that the dataset in memory
          contains only the relevant data.
Bootstrap results                                 Number of obs     =      2,272
                                                  Replications      =        250

      command:  sgini w [aweight= hours]
            G:  r(coeff)

-------------+
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
G	.3463003	.0090258	38.37	0.000	.32861	.3639906

Stratification and clustering can also be introduced in the bootstrap.

```
. gen newid = idcode
. xtset newid year
       panel variable:  newid (unbalanced)
        time variable:  year, 68 to 88, but with gaps
                delta:  1 unit
. bootstrap G=r(coeff) ///
>         , cluster(idcode) idcluster(newid) force reps(250) nodots: ///
>         sgini w [aw=hours]  if !mi(w) & year==88
Warning:  Because sgini is not an estimation command or does not set e(sample),
          bootstrap has no way to determine which observations are used in calculating
          the statistics and so assumes that all observations are used.  This means that
          no observations will be excluded from the resampling because of missing values
          or other reasons.

          If the assumption is not true, press Break, save the data, and drop the
          observations that are to be excluded.  Be sure that the dataset in memory
          contains only the relevant data.
Bootstrap results                                 Number of obs     =      2,272
                                                  Replications      =        250
      command:  sgini w [aweight= hours]
            G:  r(coeff)
                                 (Replications based on 2,272 clusters in idcode)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
G	.3463003	.0088885	38.96	0.000	.328879	.3637215

As shown in Section 4, `sgini` can be used more generally as a building block to analyze a variety of aspects of the income distribution, beyond standard inequality measures.

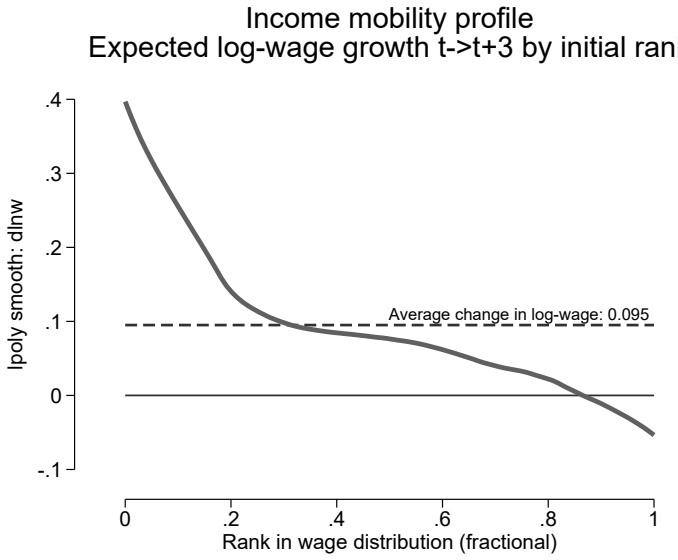
As a first example, we calculate here the bi-polarization measures described in Section 4. It requires calculating median income to split the population in two equally-sized groups, total and half-population mean incomes, and inequality indices within and between half-populations. These estimates are then combined to obtain polarization measures.

```
. preserve
. keep if year==88
(26,262 observations deleted)
. qui summarize w  , detail
. local mean = r(mean)
. local med  = r(p50)
. qui sgini w
. local sgini = r(coeff)
. su w if w<‘med’ , meanonly
. local mup = r(mean)
. su w if w>=‘med’ , meanonly
. local mur = r(mean)
. qui sgini w if w<‘med’
```

```
. local sginip = r(coeff)
. qui sgini w if w>=‘med’
. local sginir = r(coeff)
. scalar Within = 0.25 * (1/‘mean’) * (‘mup’*‘sginip’  +  ‘mur’*‘sginir’)
. scalar Between = 0.25 * (1/‘mean’) * (‘mur’  -  ‘mup’)
. scalar P1 = ( Between - Within ) / ‘sgini’
. scalar P2 = ( Between - Within ) * ‘mean’ / ‘med’
. scalar P3 =  Between / Within
. di "Within half-populations inequality: "  _col(42) %4.3f Within
Within half-populations inequality:      0.121
. di "Between half-populations inequality: " _col(42) %4.3f Between
Between half-populations inequality:     0.234
. di "Bipolarization index 1 (Silber et al.): " _col(42) %4.3f P1
Bipolarization index 1 (Silber et al.):  0.318
. di "Bipolarizarion index 2 (Wolfson): " _col(42) %4.3f P2
Bipolarizarion index 2 (Wolfson):        0.145
. di "Bipolarization index 3 (Kanbur & Zhang) : "_col(42) %4.3f P3
Bipolarization index 3 (Kanbur & Zhang) : 1.934
. restore
```

As a second example, we calculate the aggregate measures of ‘progressivity-adjusted’ wage growth developed in Jenkins & Van Kerm (2006). This is direct application of concentration coefficients. We first calculate the change in log-wage between $\mathtt{t}$ an $\mathtt{t}+3$ and then calculate its concentration coefficient with respect to initial period $\mathtt{t}$ wage. Note that the concentration with parameter $\upsilon = 1$ is the average log wage growth. The pattern can be represented by plotting a local polynomial smooth of individual log wage growth against initial fractional rank (as produced by `fracrank`). The curve illustrates regression to the mean in wages with wages growing faster at the bottom than at the top.

```
. generate dlnw = ln(F3.w) - ln(w)
(17,563 missing values generated)
. summarize dlnw
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dlnw |     10,971     .095051    .3615585  -2.524581   3.163316
. loc mn =r(mean)
. sgini dlnw , parameter(1 2 3) sortvar(w) aggregate
Generalized Concentration coefficient for dlnw against w
Note: dlnw has 3687 negative observations (used in calculations).
------------------------------------
    Variable |      v=1      v=2      v=3
-------------+----------------------
        dlnw |   0.0951   0.1526   0.1902
------------------------------------
. fracrank w , gen(prank)
. range atp 0 1 100
(28,434 missing values generated)
. label variable atp "Rank in wage distribution (fractional)"
. lpoly dlnw prank , bw(0.08) gen(profile) at(atp) nograph
```



Citation, liability, conditions of use

`sgini` is not an official Stata command. It is a free contribution to the research community. The program should work as described, but it is freely offered 'as-is'. Use at your own risk! Bug reports as well as comments and suggestions can be sent to `philippe.vankerm@liser.lu`.

Please cite as:

> Van Kerm, P. (2009), '`sgini` – Generalized Gini and Concentration coefficients (with factor decomposition) in Stata', v2.0 (revised April 2020), Luxembourg Institute of Socio-Economic Research, Esch/Alzette, Luxembourg.

References

Berger, Y. G. (2008), 'A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient', *Journal of Official Statistics*, **24**(4):541–555.

Blackorby, C. & Donaldson, D. (1980), 'A theoretical treatment of indices of absolute inequality', *International Economic Review*, **21**(1):107–36.

Chen, Z. & Roy, K. (2009), 'Calculating concentration index with repetitive values of indicators of economic welfare', *Journal of Health Economics*, **28**(1):169–175.

Chotikapanich, D. C. & Griffiths, W. (2001), 'On calculation of the extended Gini coefficient', *Review of Income and Wealth*, **47**:541–547.

Cox, N. J. (2002), 'Calculating percentile ranks or plotting positions', in *Stata FAQs*, StataCorp LP, College Station, TX, available at `http://www.stata.com/support/faqs/`.

Davidson, R. (2009), 'Reliable inference for the Gini index', *Journal of Econometrics*, **150**(1):30–40.

Demuynck, T. & van de Gaer, D. (2012), 'Inequality adjusted income growth', *Economica*, **79**(316):747–765, doi:10.1111/j.1468-0335.2012.00931.x.

Donaldson, D. & Weymark, J. A. (1980), 'A single-parameter generalization of the Gini indices of inequality', *Journal of Economic Theory*, **22**:67–86.

— (1983), 'Ethically flexible Gini indices for income distributions in the continuum', *Journal of Economic Theory*, **29**:353–358.

Fei, J. C. H., Ranis, G. & Kuo, S. W. Y. (1978), 'Growth and the family distribution of income by factor components', *Quarterly Journal of Economics*, **92**(1):17–53.

Fusco, A. & Van Kerm, P. (2020), '`bipolar`: Stata module to calculate four measures of income bipolarization', Statistical Software Components S458777, Boston College Department of Economics, `http://ideas.repec.org/c/boc/bocode/s458777.html`.

Jenkins, S. P. (1988), 'Calculating income distribution indices from micro-data', *National Tax Journal*, **41**(1):139–142.

Jenkins, S. P. & Van Kerm, P. (2006), 'Trends in income inequality, pro-poor income growth and income mobility', *Oxford Economic Papers*, **58**(3):531–48.

— (2009), '`dsginideco`: Decomposition of inequality change into pro-poor growth and mobility components', Statistical Software Components S457009, Boston College Department of Economics, `http://ideas.repec.org/c/boc/bocode/s457009.html`.

— (2016), 'Assessing individual income growth', *Economica*, **83**(332):679–703, doi:10.1111/ecca.12205.

Kakwani, N. C. (1977a), 'Applications of Lorenz curves in economic analysis', *Econometrica*, **45**(3):719–728.

— (1977b), 'Measurement of tax progressivity: an international comparison', *The Economic Journal*, **87**(345):71–80.

Lambert, P. J. (2001), *The Distribution and Redistribution of Income, a Mathematical Analysis*, Manchester University Press, Manchester, UK, 3rd edn.

Lerman, R. I. & Yitzhaki, S. (1984), 'A note on the calculation and interpretation of the Gini index', *Economics Letters*, **15**:363–368.

— (1985), 'Income inequality effects by income source: A new approach and applications to the United States', *Review of Economics and Statistics*, **67**(1):151–156.

— (1989), 'Improving the accuracy of estimates of Gini coefficients', *Journal of Econometrics*, **42**:43–47.

López-Feldman, A. (2006), 'Decomposing inequality and obtaining marginal effects', *The Stata Journal*, **6**(1):106–111.

O'Neill, D. & Van Kerm, P. (2008), 'An integrated framework for analysing income convergence', *The Manchester School*, **76**(1):1–20.

Peichl, A. & Van Kerm, P. (2007), '`progres`: Module to measure distributive effects of an income tax', Statistical Software Components S456867, Boston College Department of Economics, `http://ideas.repec.org/c/boc/bocode/s456867.html`.

Reynolds, M. & Smolensky, E. (1977), *Public Expenditures, Taxes, and the Distribution of Income: The United States, 1950, 1961, 1970*, Academic Press, , New York.

Schechtman, E. & Yitzhaki, S. (1987), 'A measure of association based on Gini's mean difference', *Communications in Statistics - Theory and Methods*, **16**(1):207–231.

— (1999), 'On the proper bounds of the Gini correlation', *Economics Letters*, **63**(2):133–138.

— (2003), 'A family of correlation coefficients based on the extended Gini index', *Journal of Economic Inequality*, **1**(2):129–146.

Sen, A. K. (1976), 'Real national income', *Review of Economic Studies*, **43**(1):19–39.

Silber, J., Hanoka, M. & Deutsch, J. (2007), 'On the link between the concepts of kurtosis and bipolarization', *Economics Bulletin*, **4**(36):1–6.

van Doorslaer, E., Wagstaff, A., Bleichrodt, H., Calonge, S., Gerdtham, U.-G., Gerfin, M., Geurts, J., Gross, L., Häkinnen, U., Leu, R. E., O'Donnell, O., Propper, C., Puffer, F., Rodriguez, M., Sundberg, G. & Winkelhake, O. (1997), 'Income-related inequalities in health: Some international comparisons', *Journal of Health Economics*, **16**:93–112.

Van Kerm, P. (2020), '`sgini`: Stata module to compute generalized Gini and concentration coefficients, Gini correlations and fractional ranks', Statistical Software Components S458778, Boston College Department of Economics, `http://ideas.repec.org/c/boc/bocode/s458778.html`.

Wolfson, M. C. (1994), 'When inequalities diverge', *American Economic Review*, **84**(2):353–58.

Yitzhaki, S. (1983), 'On an extension of the Gini inequality index', *International Economic Review*, **24**:617–628.

— (1998), 'More than a dozen alternative ways of spelling Gini', in D. J. Slottje (ed.), *Research on Economic Inequality*, vol. 8, JAI Press, Elsevier Science, 13–30.

Yitzhaki, S. & Schechtman, E. (2005), 'The properties of the extended Gini measures of variability and inequality', *METRON International Journal of Statistics*, **63**(3):401–433.

Yitzhaki, S. & Wodon, Q. (2004), 'Inequality, mobility, and horizontal equity', in Y. Amiel & J. A. Bishop (eds.), *Research on Economic Inequality (*Studies on Economic Well-Being: Essays in Honor of John P. Formby*)*, vol. 12, JAI Press, Elsevier Science, 177–198.

Zhang, X. & Kanbur, R. (2001), 'What difference do polarisation measures make? An application to China', *Journal of Development Studies*, **37**(3):85–98.

Acknowledgements

This work was part of the MeDIM project (*Advances in the Measurement of Discrimination, Inequality and Mobility*) supported by the Luxembourg 'Fonds National de la Recherche' (contract FNR/06/15/08) and by core funding for CEPS/INSTEAD by the Ministry of Culture, Higher Education and Research of Luxembourg.