sgini

Generalized Gini and Concentration coefficients (with factor decomposition) in Stata

Philippe Van Kerm
CEPS/INSTEAD, Luxembourg‡

Abstract `sgini` is a user-written Stata package to compute generalized Gini and concentration coefficients. This note describes syntax, formulas and usage examples.

September 2009 (revised February 2010)

Keywords `sgini` ; Stata ; generalized Gini ; Concentration coefficient

JEL Classification C88; D31

1 Introduction

This document describes `sgini`, a lightweight command for calculations of generalized Gini (a.k.a. S-Gini) and Concentration coefficients from unit-record data (not grouped data) in Stata. `sgini` computes 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 optionally report 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 economic applications.

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

```
net install sgini , from(http://medim.ceps.lu/stata)
```

2 Generalized Gini and Concentration coefficients

2.1 Covariance-based formulations

The Gini coefficient is one of 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]

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

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

‡Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques/International Networks for Studies in Technology, Environment, Alternatives, Development. B.P. 48, L-4501 Differdange, Luxembourg. `http://www.ceps.lu`. E-mail: `philippe.vankerm@ceps.lu`.

[1] The latest version of the `sgini` package is 1.1.0 (of 2010-02-05). Stata 8.2 or later is required.

[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.

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

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. A generalized Concentration coefficient can be similarly defined as

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

Note finally that while measures of inequality are typically taken as relative – one considers the distribution of, say, income shares, 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

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

while the second case is obtained as

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

2.2 Calculation

Covariance-based expressions for the generalized Gini and Concentration coefficients are particularly 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.

Attention needs to be paid to handle carefully tied values in the computation of the fractional ranks. Consider a sample of N observations on a variable Y with associated sample 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:

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

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

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

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 coefficient estimate independent on the sample/population size.[3]

3 Applications

Gini coefficients are popular measures of inequality by themselves. 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. But 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.

3.1 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

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

where $\mathrm{CONC}(Y^k, Y;\upsilon)$ 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 $\mathrm{CONC}(Y^k, Y;\upsilon)$ can further be expressed as

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

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

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

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:

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

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

[3]See Yitzhaki & Schechtman (2005), Berger (2008), and Davidson (2009) for recent discussions of this issue. Also see Cox (2002) for a more general discussion of (fractional) ranks. In Stata, `cumul` or `glcurve`'s `pvar()` option are sometimes used to estimate ranks that are plugged in formula for the Gini or Concentration index. Using these estimates of the empirical distribution function is not advisable here because ranks estimated in this way are not fractional and the resulting estimate of the Gini is therefore not population invariant (and so depend on the sample size).

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

3.2 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

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

where $R = (\text{CONC}(X^{\text{post}}, X^{\text{pre}}) - \text{GINI}(X^{\text{post}}))$ 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 number of Gini and Concentration coefficients.

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

3.3 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:

$$\Delta(\upsilon) = R(\upsilon) - P(\upsilon)$$

where

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

and

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

$P(\upsilon)$ 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. $R(\upsilon)$ 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 $P(\upsilon)$ and $R(\upsilon)$. Note that $R(\upsilon)$ 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 $P(\upsilon)$ 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, 2009b).

[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*).

3.4 Income polarization

The Gini coefficient has also been used as a building block of several 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 M(Y), mean income $\mu(\mathsf{Y})$, mean income among the poor $\mu(\mathsf{Y}^\mathsf{P})$, mean income among the rich $\mu(\mathsf{Y}^\mathsf{R})$, the Gini coefficient among the poor $\mathrm{GINI}(\mathsf{Y}^\mathsf{P})$, the Gini coefficient among the rich $\mathrm{GINI}(\mathsf{Y}^\mathsf{R})$ and the overall Gini, 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:

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

and

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

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

$$\mathsf{P}_1 = \frac{\mathrm{Between}(\mathsf{Y}^\mathsf{P},\mathsf{Y}^\mathsf{R}) - \mathrm{Within}(\mathsf{Y}^\mathsf{P},\mathsf{Y}^\mathsf{R})}{\mathrm{GINI}(\mathsf{Y})},$$

the index of Wolfson (1994) as

$$\mathsf{P}_2 = \left(\mathrm{Between}(\mathsf{Y}^\mathsf{P},\mathsf{Y}^\mathsf{R}) - \mathrm{Within}(\mathsf{Y}^\mathsf{P},\mathsf{Y}^\mathsf{R})\right)\frac{\mu(\mathsf{Y})}{\mathrm{Med}(\mathsf{Y})},$$

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

$$\mathsf{P}_3 = \frac{\mathrm{Between}(\mathsf{Y}^\mathsf{P},\mathsf{Y}^\mathsf{R})}{\mathrm{Within}(\mathsf{Y}^\mathsf{P},\mathsf{Y}^\mathsf{R})}.$$

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

4 The `sgini` command

`sgini` is a lightweight command to compute generalized Gini and concentration coefficents from unit-record data in Stata using the formulas given in Section 2. `sgini` can also report the decomposition by income source (a.k.a. a factor decomposition) if requested. `sgini` comes with a companion command `fracrank` for generating fractional ranks as *infra*.

4.1 Syntax

The syntax of `sgini` is as follows:

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

`fweight`, `aweight`, `pweight` and `iweight` 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 param and with ordering variable of option sortvar. Decomposition by income source is optionally computed. See detailed option descriptions below.

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

The syntax of the accompanying fracrank is

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

fweight and aweight are allowed. 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.

4.2 Options

4.2.1 sgini options

param(*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.

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).

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.

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

4.2.2 fracrank options

generate(*newvarname*) specifies the name for the created variable.

4.3 Saved 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

4.4 Dependencies on user-written packages

`sgini` does not require other user-written packages.

5 Example

This example illustrates 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 `tsset` the data as appropriate.

```
. cap use http://www.stata-press.com/data/r9/nlswork , clear

. tsset 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 |    0.273

. sgini w , param(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 |    1.121    1.652    1.981    2.211    2.384    2.521
```

The third example illustrates usage of multiple variables in *varlist* (and of time-series operators) and estimation of 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 |    0.202
         L.w |    0.192
        L2.w |    0.193

. 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 |    0.202    0.284
         L.w |    0.158    0.220
        L2.w |    0.139    0.192

. return list
scalars:
              r(sum_w) =  3481
                  r(N) =  3481
              r(coeff) =  .2023171629989085
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:
                           L.         L2.
               w           w           w           w
Coeff  .20231716    .1581236   .13867147   .28359513

          param2:     param2:
               L.         L2.
               w           w
```

```
Coeff  .21983376  .19212248
```

Specifying option `sourcedecomposition` produces the factor decomposition.

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

--------------------------
    Variable |        v=2
-------------+------------
           w |      0.202
         L.w |      0.192
        L2.w |      0.193
--------------------------

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

Parameter: v=2
------------------------------------------------------------------------------
    Variable |         s         g         r       g*r     s*g*r  s*g*r/G s*g*r/G-s
-------------+----------------------------------------------------------------
           w |     0.351     0.202     0.929     0.188     0.066     0.366     0.015
         L.w |     0.335     0.192     0.939     0.180     0.060     0.335    -0.000
        L2.w |     0.314     0.193     0.892     0.172     0.054     0.299    -0.015
-------------+----------------------------------------------------------------
       TOTAL |     1.000     0.180     1.000     0.180     0.180     1.000     0.000
------------------------------------------------------------------------------

. return list

scalars:
              r(sum_w) =  3481
                  r(N) =  3481
              r(coeff) =  .2023171629989085

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

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

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

```
. gen newid = idcode

. tsset 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) reps(250) nodots: ///
>    sgini w if !mi(w)

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      =     28534
                                                Replications       =       250

      command:  sgini w
            G:  r(coeff)
                               (Replications based on 4711 clusters in idcode)
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           G |   .2731835   .0029554    92.44   0.000     .2673911     .278976
------------------------------------------------------------------------------

. jackknife G=r(coeff) ///
>    , cluster(idcode) idcluster(newid) rclass nodots:    ///
>    sgini w if !mi(w)
Jackknife results                               Number of obs      =     28534
                                                Replications       =      4711

      command:  sgini w if !mi(w)
            G:  r(coeff)
          n():  r(N)
                               (Replications based on 4711 clusters in idcode)
------------------------------------------------------------------------------
             |              Jackknife
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           G |   .2731835   .0029859    91.49   0.000     .2673298    .2790373
------------------------------------------------------------------------------
```

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@ceps.lu`.

Please cite as:

> Van Kerm, P. (2009), '`sgini` – Generalized Gini and Concentration coefficients (with factor decomposition) in Stata', v1.1 (revised February 2010), CEPS/INSTEAD, Differdange, 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.

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*, available online – doi:10.1016/j.jeconom.2008.11.004.

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.

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`.

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.

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.

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 is 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.