UNIVERSITÄT
BERN

Faculty of Business, Economics and Social Sciences

Department of Social Sciences

University of Bern Social Sciences Working Paper No. 15

Estimating Lorenz and concentration curves in Stata

Ben Jann

This paper is forthcoming in the Stata Journal.

Current version: October 27, 2016
First version: January 12, 2016

http://ideas.repec.org/p/bss/wpaper/15.html
http://econpapers.repec.org/paper/bsswpaper/15.htm

University of Bern
Department of Social Sciences
Fabrikstrasse 8
CH-3012 Bern

Tel. +41 (0)31 631 48 11
Fax +41 (0)31 631 48 17
info@sowi.unibe.ch
www.sowi.unibe.ch

Estimating Lorenz and concentration curves in Stata

Ben Jann
Institute of Sociology
University of Bern
ben.jann@soz.unibe.ch

October 27, 2016

Abstract

Lorenz and concentration curves are widely used tools in inequality research. In this paper I present a new Stata command called `lorenz` that estimates Lorenz and concentration curves from individual-level data and, optionally, displays the results in a graph. The `lorenz` command supports relative as well as generalized, absolute, unnormalized, or custom-normalized Lorenz or concentration curves, and provides tools for computing contrasts between different subpopulations or outcome variables. Variance estimation for complex samples is fully supported.

Keywords: Stata, `lorenz`, Lorenz curve, concentration curve, inequality, income distribution, wealth distribution, graphics

Contents

1 Introduction

Lorenz curves and concentration curves are widely used tools for the analysis of economic inequality and redistribution (see, e.g., Cowell, 2011; Lambert, 2001). Yet, no official command for the estimation of Lorenz and concentration curves is offered in Stata.

In this paper I present an implementation of such a command, called `lorenz`. Although other user commands with related functionality do exist,[1] I believe that `lorenz` is a worthwhile contribution that will prove beneficial to inequality researchers. The command computes and, optionally, graphs relative, total (unnormalized), generalized, or absolute Lorenz and concentration curves from individual-level data. Standard errors and confidence intervals are provided and estimation from complex samples is fully supported. Furthermore, `lorenz` is well suited for subpopulation analysis and offers options to compute contrasts between subpopulations or between outcome variables (including standard errors). The command also offers custom normalization of results, which can be useful for comparing subpopulations or outcome variables. Finally, `lorenz` saves its results in the `e()` returns for easy processing by post-estimation commands.

In the remainder of the paper I will first discuss the relevant methods and formulas and then present the syntax and options of the `lorenz` command. After that, the usage of the command will be illustrated by a number of examples.

[1] Examples are `glcurve` (Jenkins and Van Kerm, 1999; Van Kerm and Jenkins, 2001), `svylorenz` (Jenkins, 2006), `clorenz` (Abdelkrim, 2005), or `alorenz` (Azevedo and Franco, 2006).

2 Methods and formulas

2.1 Lorenz curve

Let X be the outcome variable of interest (e.g. income). The cumulative distribution function of X is given as $F_X(x) = \Pr\{X \leq x\}$ and the quantile function (the inverse of the distribution function) is given as $Q_X(p) = F_X^{-1}(p) = \inf\{x|F_X(x) \geq p\}$ with $p \in [0, 1]$. For continuous X, the ordinates of the *relative Lorenz curve* are given as

$$L_X(p) = \frac{\int_{-\infty}^{Q_X^p} y \, dF_X(x)}{\int_{-\infty}^{\infty} x \, dF_X(x)}$$

(see, e.g., Cowell, 2000; Lambert, 2001; Hao and Naiman, 2010). Intuitively, a point on the Lorenz curve quantifies the proportion of total outcome of the poorest $p \cdot 100$ percent of the population. This can easily be seen in the finite population form of $L_X(p)$, which is given as

$$L_X(p) = \frac{\sum_{i=1}^{N} X_i I\{X_i \leq Q_X^p\}}{\sum_{i=1}^{N} X_i}$$

with $I\{A\}$ as an indicator function being equal to 1 if A is true and 0 else.

Furthermore, let J_i be an indicator for whether observation i belongs to subpopulation j or not (i.e., $J_i = 1$ if observation i belongs to subpopulation j and $J_i = 0$ else). The finite population form of the Lorenz curve of X in subpopulation j can then be written as

$$L_X^j(p) = \frac{\sum_{i=1}^{N} X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\sum_{i=1}^{N} X_i J_i}$$

where $Q_X^{p,j}$ is the p-quantile of X in subpopulation j. The population-wide Lorenz curve is obtained by setting $J_i = 1$ for all observations.

Lorenz curves are typically displayed graphically with p on the horizontal axis and $L_X(p)$ on the vertical axis, although Lorenz (1905) originally proposed an opposite layout.

2.2 Equality gap curve

The *equality gap curve* quantifies the degree to which the proportion of total outcome of the poorest $p \cdot 100$ percent of the population deviates from the proportion of total outcome these population members would get under an equal distribution. That is, the equality gap curve is equal to the difference between the equal distribution diagonal and the Lorenz curve. Formally, the (finite population form of the) equality gap curve of Y in subpopulation j is given as

$$EG_X^j(p) = p - \frac{\sum_{i=1}^{N} X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\sum_{i=1}^{N} X_i J_i} = p - L_X^j(p)$$

$EG_X(p)$ is equal to the proportion of total outcome that would have to be relocated to the poorest $p \cdot 100$ percent in order to provide them an average outcome equal to the population average.

2.3 Total (unnormalized) Lorenz curve

In the finite population, define the (subpopulation specific) *total Lorenz curve* as

$$TL_X^j(p) = \sum_{i=1}^{N} X_i I\{X_i \leq Q_X^{p,j}\} J_i$$

The total Lorenz curve quantifies the cumulative sum of outcomes among the poorest $p \cdot 100$ percent of the (sub-)population.

2.4 Generalized Lorenz curve

The ordinates of the (relative) Lorenz curve refer to cumulative outcome proportions. Hence, $L_X(1) = 1$. In contrast, the ordinates of the *generalized Lorenz curve*, $GL_X(p)$, refer to the cumulative outcome *average*. Hence $GL_X(1) = \overline{X}$, where $\overline{X}$ is the mean of X. Formally, the generalized Lorenz curve can be defined as

$$GL_X(p) = \int_{-\infty}^{Q_X^p} x \, dF_X(x)$$

the finite population form of which is

$$GL_X(p) = \frac{1}{N} \sum_{i=1}^{N} X_i I\{X_i \leq Q_X^p\}$$

(see, e.g., Shorrocks, 1983; Cowell, 2000; Lambert, 2001). Furthermore, for subpopulation j, the generalized Lorenz curve can be written as

$$GL_X^j(p) = \frac{1}{\sum_{i=1}^{N} J_i} \sum_{i=1}^{N} X_i I\{X_i \leq Q_X^{p,j}\} J_i$$

where $\sum_{i=1}^{N} J_i$ is equal to the subpopulation size.

2.5 Absolute Lorenz curve

The *absolute Lorenz curve* quantifies the degree to which the generalized Lorenz curve deviates from the equal distribution line in terms of the cumulative outcome average (see, e.g.,

Moyes, 1987). Formally, the (finite population form of the) absolute Lorenz curve of Y in subpopulation j is given as

$$AL_X^j(p) = \frac{1}{\sum_{i=1}^N J_i}\left(\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\}J_i - p\sum_{i=1}^N X_i J_i\right) = GL_X^j(p) - p\frac{\sum_{i=1}^N X_i J_i}{\sum_{i=1}^N J_i}$$

2.6 Concentration curve

The Lorenz curve of outcome variable X refers cumulative outcome proportions of population members ranked by the values of X. Using an alternative ranking variable Y, while still measuring outcome in terms of X, leads to the so-called *concentration curve*. Formally, the (relative) concentration curve of X with respect to Y can be defined as

$$L_{XY}(p) = \frac{\int_{-\infty}^{Q_Y^p}\int_{-\infty}^{\infty} x f_{XY}(x,y)\,dx\,dy}{\int_{-\infty}^{\infty} x\,dF_X(x)}$$

where Q_Y^p is the p-quantile of the distribution of Y and $f_{XY}(x,y)$ is the density of the joint distribution of X and Y (see, e.g., Bishop et al., 1994). In the finite population the concentration curve simplifies to

$$L_{XY}(p) = \frac{\sum_{i=1}^N X_i I\{Y_i \leq Q_Y^p\}}{\sum_{i=1}^N X_i}$$

Furthermore, for subpopulation j the concentration curve can be written as

$$L_{XY}^j(p) = \frac{\sum_{i=1}^N X_i I\{Y_i \leq Q_Y^{p,j}\}J_i}{\sum_{i=1}^N X_i J_i}$$

Total, generalized, or absolute concentration curves can be defined analogously.

2.7 Renormalization

Relative Lorenz curves are normalized with respect to the total of the analyzed outcome variable in the given population or subpopulation. Depending on context, it may be useful to apply a different type of normalization. For example, when analyzing labor income, we may want to express results with respect total income (labor income plus capital income). Likewise, when analyzing a subpopulation, we may be interested in results relative to another subpopulation or relative to the overall population.

To normalize the relative Lorenz curve or the equality gap curve of X with respect to the total of Z (where Z may be the sum of several variables, possibly including X), let

$$L_X^{j,Z}(p) = \frac{\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\}J_i}{\sum_{i=1}^N Z_i J_i} \qquad \text{and} \qquad EG_X^{j,Z}(p) = p - L_X^{j,Z}(p)$$

Likewise, to normalize with respect to a fixed (subpopulation) total τ, let

$$L_X^{j,\tau}(p) = \frac{\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\tau} \qquad \text{and} \qquad EG_X^{j,\tau}(p) = p - L_X^{j,\tau}(p)$$

To normalize the Lorenz curve of subpopulation j with respect to the total in subpopulation r (where subpopulation r may include subpopulation j), define

$$L_X^{jr}(p) = \frac{\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\sum_{i=1}^N X_i R_i}$$

where R_i is an indicator for whether observation i belongs to subpopulation r or not. For example, if r is the entire population (including subpopulation j), then $L_X^{j,r}(p)$ is the proportion of the population-wide outcome that goes to the poorest $p \cdot 100$ percent of subpopulation j. In contrast, since the equality gap curve is supposed to quantify the deviation from the equal distribution line, the renormalized equality gap curve of subpopulation j with respect to the total in subpopulation r should be defined as

$$EG_X^{jr}(p) = p\frac{\sum_{i=1}^N J_i}{\sum_{i=1}^N R_i} - L_X^{jr}(p)$$

where $p \sum_{i=1}^N J_i / \sum_{i=1}^N R_i$ is the outcome share of the poorest $p \cdot 100$ percent of subpopulation j if all population members would receive the same outcome.

The normalized Lorenz curve ordinates $L_X^{jr}(p)$ express outcome shares in subpopulation j relative to the total outcome of subpopulation r. An alternative is to renormalize Lorenz curve ordinates in a way such that they are relative to the total that would be observed in subpopulation j, if all members of subpopulation j would receive the average outcome of subpopulation r. This can be achieved by rescaling the total by relative group sizes, that is,

$$L_X^{j\overline{r}}(p) = \frac{\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\frac{\sum_{i=1}^N J_i}{\sum_{i=1}^N R_i} \sum_{i=1}^N X_i R_i} \qquad \text{and} \qquad EG_X^{j\overline{r}}(p) = p - L_X^{j\overline{r}}(p)$$

Note that there is a close relation between $L_X^{j\overline{r}}(p)$ and generalized Lorenz curves: The ratio of $L_X^{j\overline{r}}(p)$ from two subpopulations is equal to the ratio of the generalized Lorenz curves from these subpopulations.

Combining normalization with respect to a different subpopulation and normalization with respect to the total of a different outcome variable or a fixed total leads to

$$L_X^{jr,Z}(p) = \frac{\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\sum_{i=1}^N Z_i R_i} \qquad EG_X^{jr,Z}(p) = p\frac{\sum_{i=1}^N J_i}{\sum_{i=1}^N R_i} - L_X^{jr,Z}(p)$$

$$L_X^{j\overline{r},Z}(p) = \frac{\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\frac{\sum_{i=1}^N J_i}{\sum_{i=1}^N R_i} \sum_{i=1}^N Z_i R_i} \qquad EG_X^{j\overline{r},Z}(p) = p - L_X^{j\overline{r},Z}(p)$$

and

$$L_X^{jr,\tau}(p) = \frac{\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\tau} \qquad EG_X^{jr,\tau}(p) = p\frac{\sum_{i=1}^N J_i}{\sum_{i=1}^N R_i} - L_X^{jr,\tau}(p)$$

$$L_X^{j\overline{r},\tau}(p) = \frac{\sum_{i=1}^N X_i I\{X_i \leq Q_X^{p,j}\} J_i}{\frac{\sum_{i=1}^N J_i}{\sum_{i=1}^N R_i}\tau} \qquad EG_X^{j\overline{r},\tau}(p) = p - L_X^{j\overline{r},\tau}(p)$$

Analogous renormalizations can be applied to concentration curves. Simply replace $I\{X_i \leq Q_X^{p,j}\}$ in the above formulas by $I\{Y_i \leq Q_Y^{p,j}\}$.

2.8 Contrasts

To analyze distributional differences is helpful to compute contrasts between Lorenz curves. For example, the difference

$$L_X(p) - L_Y(p)$$

may be used to evaluate whether distribution X Lorenz dominates distribution Y. Likewise, the difference

$$GL_X(p) - GL_Y(p)$$

may be used to evaluate whether distribution X generalized Lorenz dominates distribution Y. Dominance is given if the difference is positive for all p. As shown by Atkinson (1970), if distribution X Lorenz dominates distribution Y then distribution X can be seen as less unequal than distribution Y under weak conditions. Likewise, if distribution X generalized Lorenz dominates distribution Y then distribution X can be seen as preferable over distribution Y in terms of welfare under weak conditions (see, e.g., Lambert, 2001).

Depending on context, it may also be practical to define contrasts as ratios, that is, $L_X(p)/L_Y(p)$, or as logarithms of ratios, that is, $\ln(L_X(p)/L_Y(p))$.

2.9 Point estimation

Given is a sample X_i, $i = 1, \ldots, n$, with sampling weights w_i. Furthermore, let subscripts in parentheses refer to observations sorted in ascending order of X. $L_X(p)$ can then be estimated as

$$\widehat{L}_X(p) = (1-\gamma)\widetilde{X}_{i_p-1} + \gamma\widetilde{X}_{i_p},$$

where

$$\gamma = \frac{p - \widehat{p}_{i_p-1}}{\widehat{p}_{i_p} - \widehat{p}_{i_p-1}}, \quad \widetilde{X}_{i_p} = \frac{\sum_{i=1}^{i_p} w_{(i)} X_{(i)}}{\sum_{i=1}^{n} w_i X_i}, \quad \text{and} \quad \widehat{p}_{i_p} = \frac{\sum_{i=1}^{i_p} w_{(i)}}{\sum_{i=1}^{n} w_i}$$

and where i_p is set such that $\widehat{p}_{i_p-1} < p \leq \widehat{p}_{i_p}$. Using this approach, ties in X are broken proportionally and linear interpolation (corresponding to quantile definition 4 in Hyndman

and Fan, 1996) is applied where the distribution function of X is flat. An alternative would be to avoid linear interpolation and estimate $L_X(p)$ as

$$\widehat{L}_X(p) = \widetilde{X}_{i_p} = \frac{\sum_{i=1}^{i_p} w_{(i)} X_{(i)}}{\sum_{i=1}^{n} w_i X_i}$$

(corresponding to quantile definition 1 in Hyndman and Fan, 1996). The first approach appears preferable over the second approach, because the latter requires arbitrary decisions on the sort order within ties of X to obtain stable results in presence of sampling weights.

Analogous formulas can be used to estimate total, generalized, absolute or renormalized Lorenz curves. For concentration curves, the observations are sorted in order of Y instead of X and, to enforce stable results, X values can be averaged within ties of Y.

2.10 Variance estimation

Following Binder and Kovacevic (1995) and Kovačević and Binder (1997), approximate variance estimates for Lorenz ordinates can be obtained by estimating the total (see [R] **total**)—possibly accounting for complex survey design (see [SVY] **svy estimation**)—of residual variables defined as

$$u_i^j(p) = \frac{\left((X_i - \widehat{Q}_X^{p,j}) I\{X_i \leq \widehat{Q}_X^{p,j}\} + p\, \widehat{Q}_X^{p,j}\right) J_i - a_i}{b}$$

where a_i and b are as described in table 1 (for details see Jann, 2016).[2] For concentration curves, replace $I\{X_i \leq \widehat{Q}_X^{p,j}\}$ by $I\{Y_i \leq \widehat{Q}_Y^{p,j}\}$ and replace $\widehat{Q}_X^{p,j}$ by $\widehat{E}(X|Y = Q_Y^{p,j}, J = 1)$.[3] Furthermore, variance estimates for contrasts can be obtained by the delta method as outlined in Jann (2016).

[2] When computing the u variables, the `lorenz` command presented below uses definition 4 in Hyndman and Fan (1996) to determine $\widehat{Q}_X^{p,j}$ (or definition 1, depending on the method used for estimating the Lorenz ordinates). Furthermore, in analogy to the approach employed for point estimation, ties in X are broken when determining $I\{X_i \leq \widehat{Q}_X^{p,j}\}$ (based on observations sorted by w_i within ties, which is an arbitrary decision to enforce stable results). Depending on sample design, terms $\frac{1}{w_i n}$ and $\frac{\tau}{w_i n}$ in the formulas in table 1 require modification; an alternative is to simply set these terms to zero (see Jann, 2016). Finally, for equality gap curves, use $-u$ instead of u.

[3] In the `lorenz` command presented below, $E(X|Y = Q_Y^{p,j}, J = 1)$, the expected value of X at the p-quantile of Y in subpopulation j, is estimated by local linear regression using the Epanechnikov kernel and the default rule-of-thumb bandwidth as described in [R] **lpoly**.

Table 1: Definitions of a_i and b

	a_i	b
$L_X^j(p), EG_X^j(p)$	$X_i J_i \widehat{L}_X^j(p)$	$\sum_i w_i X_i J_i$
$TL_X^j(p)$	$\frac{1}{w_i n} \widehat{TL}_X^j(p)$	1
$GL_X^j(p)$	$J_i \widehat{GL}_X^j(p)$	$\sum_i w_i J_i$
$AL_X^j(p)$	$(\widehat{AL}_X^j(p) + pX_i)J_i$	$\sum_i w_i J_i$
$L_X^{j,Z}(p), EG_X^{j,Z}(p)$	$Z_i J_i \widehat{L}_X^{j,Z}(p)$	$\sum_i w_i Z_i J_i$
$L_X^{j,\tau}(p), EG_X^{j,\tau}(p)$	$\frac{\tau}{w_i n} \widehat{L}_X^{j,\tau}(p)$	τ
$L_X^{jr}(p)$	$X_i R_i \widehat{L}_X^{jr}(p)$	$\sum_i w_i X_i R_i$
$EG_X^{jr}(p)$	$\left(\frac{\sum_k w_k X_k R_k}{\sum_k w_k J_k} J_i - \frac{\sum_k w_k X_k R_k}{\sum_k w_k R_k} R_i\right) p \frac{\sum_k w_k J_k}{\sum_k w_k R_k} + X_i R_i \widehat{L}_X^{jr}(p)$	$\sum_i w_i X_i R_i$
$L_X^{j\overline{r}}(p), EG_X^{j\overline{r}}(p)$	$\left(X_i R_i - \frac{\sum_k w_k X_k R_k}{\sum_k w_k R_k} R_i + \frac{\sum_k w_k X_k R_k}{\sum_k w_k J_k} J_i\right) \times \frac{\sum_k w_k J_k}{\sum_k w_k R_k} \widehat{L}_X^{j\overline{r}}(p)$	$\frac{\sum_i w_i J_i}{\sum_i w_i R_i} \sum_i w_i X_i R_i$
$L_X^{jr,\tau}(p)$	$\frac{\tau}{w_i n} \widehat{L}_X^{jr,\tau}(p)$	τ
$EG_X^{jr,\tau}(p)$	$\left(\frac{\tau J_i}{\sum_k w_k J_k} - \frac{\tau R_i}{\sum_k w_k R_k}\right) p \frac{\sum_k w_k J_k}{\sum_k w_k R_k} + \frac{\tau}{w_i n} \widehat{L}_X^{jr,\tau}(p)$	τ
$L_X^{j\overline{r},\tau}(p), EG_X^{j\overline{r}}(p)$	$\left(\frac{\tau}{w_i n} - \frac{\tau R_i}{\sum_k w_k R_k} + \frac{\tau J_i}{\sum_k w_k J_k}\right) \frac{\sum_k w_k J_k}{\sum_k w_k R_k} \widehat{L}_X^{j\overline{r},\tau}(p)$	$\frac{\sum_i w_i J_i}{\sum_i w_i R_i} \tau$

(all sums are across the entire sample)

3 The lorenz command

Three subcommands are provided. `lorenz estimate` computes the Lorenz curve ordinates and their variance matrix; `lorenz contrast` computes differences in Lorenz curve ordinates between outcome variables or subpopulations based on the results by `lorenz estimate`; and `lorenz graph` draws a line graph from the results provided by `lorenz estimate` or `lorenz contrast`.

To install `lorenz`, type

```
. ssc install lorenz
```

3.1 Syntax of lorenz estimate

The syntax of `lorenz estimate` is

`lorenz` [`estimate`] *varlist* [*if*] [*in*] [*weight*] [`,` *options*]

where `pweights`, `iweights`, and `fweights` are allowed; see [U] **11.1.6 weight**. For each specified variable, Lorenz curve ordinates are tabulated along with their standard errors and confidence intervals.[4] Only one variable is allowed in *varlist*, if the `over()` option is specified (see below). `lorenz` assumes subcommand `estimate` as the default; typing the word "`estimate`" is only required in case of a name conflict between the first element in *varlist* and the other subcommands of `lorenz` (see below). Options are as follows.

Main

`gap` to compute equality gap curves instead of relative Lorenz curves.

`sum` to compute total (unnormalized) Lorenz curves instead of relative Lorenz curves.

`generalized` to compute generalized Lorenz curves instead of relative Lorenz curves.

`absolute` to compute absolute Lorenz curves instead of relative Lorenz curves.

Only one of `gap`, `sum`, `generalized`, and `absolute` is allowed.

`percent` to express results as percentages instead of proportions. `percent` is not allowed in combination with `sum`, `generalized`, or `absolute`.

`normalize(`*spec*`)` to normalize Lorenz ordinates with respect to the specified total (not allowed in combination with `sum`, `generalized`, or `absolute`). *spec* is

[*over*`:`][*total*][`,` `average`]

[4] Variance estimation is not supported for `iweights` and `fweights`. To compute standard errors and confidence intervals in case of `fweights`, apply `lorenz` to the expanded data (see [R] **expand**).

where *over* may be

`.`	the subpopulation at hand (the default)
#	the subpopulation identified by value *#*
`#`*#*	the *#*th subpopulation
`total`	the total across all subpopulations

and *total* may be

`.`	the total of the variable at hand (the default)
`*`	the total of the sum across all analyzed outcome variables
varlist	the total of the sum across the variables in *varlist*
#	a total equal to *#*

total specifies the variable(s) from which the total is to be computed, or sets the total to a fixed value. If multiple variables are specified, the total across all specified variables is used (*varlist* may contain external variables that are not among the list of analyzed outcome variables). *over* selects the reference population from which the total is to be computed; *over* is only allowed if the `over()` option has been specified (see below). Suboption `average` causes subpopulation sizes (sum of weights) to be taken into account so that the results are relative to the average outcome in the reference population; this is only relevant if *over* is provided.

`gini` to report the Gini coefficients or concentration indices of the distributions (for computational details see Jann, 2016).

Percentiles

`nquantiles(`*#*`)` to specify the number of (equally spaced) percentiles to be used to determine the Lorenz ordinates (plus an additional point at the origin). The default is `nquantiles(20)`. This is equivalent to typing `percentiles(0(5)100)`.

`percentiles(`*numlist*`)` to specify, as percentages, the percentiles for which the Lorenz ordinates are to be computed. The numbers in *numlist* must be within 0 and 100. Shorthand conventions as described in [U] **11.1.8 numlist** may be applied. For example, to compute Lorenz ordinates from 0 to 100% in steps of 1 percentage point, type `percentiles(0(1)100)`. The numbers provided in `percentiles()` do not need to be equally spaced and do not need to cover the whole distribution. For example, to focus on the top 10% and use an increased resolution for the top 1% you could type `percentiles(90(1)98 99(0.1)100)`.

`pvar(`*pvar*`)` to compute concentration curves with respect to variable *pvar*.

`step` to determine the Lorenz ordinates from the step function of cumulative outcomes. The default is to employ linear interpolation in regions where the step function is flat.

Over

`over(`*varname*`)` to repeat results for each subpopulation defined by the values of *varname*.

`total` to report additional overall results across all subpopulations. `total` is only allowed if `over()` is specified.

Contrast/Graph

`contrast(`*spec*`)` to compute differences in Lorenz ordinates between outcome variables or between subpopulations. *spec* is

[*base*] [`, ratio lnratio`]

where *base* is the name of the outcome variable or the value of the subpopulation to be used as base for the contrasts. If *base* is omitted, adjacent contrasts across outcome variables or subpopulations are computed (or contrasts with respect to the total if total results across subpopulations have been requested).

Use suboption `ratio` to compute contrasts as ratios or suboption `lnratio` to compute contrasts as logarithms of ratios. The default is to compute contrasts as differences.

`graph`[`(`*options*`)`] to draw a line graph of the results. *options* are as described for `lorenz graph` below.

SE/SVY

`vce(`*vcetype*`)` to determine how standard errors and confidence intervals are computed where *vcetype* may be:

`analytic`
`cluster` *clustvar*
`bootstrap` [`,` *bootstrap_ options*]
`jackknife` [`,` *jackknife_ options*]

`analytic` is the default. See [R] **bootstrap** and [R] **jackknife** for *bootstrap_ options* and *jackknife_ options*, respectively.

`svy`[`(`*subpop*`)`] for taking the survey design as set by `svyset` into account; see [SVY] **svyset**. Specify *subpop* to restrict survey estimation to a subpopulation, where *subpop* is

[*varname*] [*if*]

The subpopulation is defined by observations for which *varname* $\neq 0$ and for which the *if* condition is met. See [SVY] **subpopulation estimation** for more information on subpopulation estimation.

The `svy` option is only allowed if the variance estimation method set by `svyset` is Taylor linearization (the default). For other variance estimation methods the usual `svy` prefix

command can be used; see [SVY] **svy**. For example, type "`svy brr: lorenz ...`" to use BRR variance estimation. `lorenz` does not allow the `svy` prefix for Taylor linearization due to technical reasons. This is why the `svy` option is provided.

`nose` to suppress the computation of standard errors and confidence intervals. Use the `nose` option to speed-up computations, for example, when applying a prefix command that employs replication techniques for variance estimation, such as, e.g., [SVY] **svy jackknife**. Option `nose` is not allowed together with `vce()` or `svy`.

Reporting

`level(#)` to set the level of confidence intervals; see [R] **level**.

`noheader` to suppress the output header, `notable` to suppress the coefficient table, and `nogtable` to suppress the table containing Gini coefficients.

display_options such as `cformat()` or `coeflegend` to format the coefficient table. See [R] **estimation options**.

3.2 Syntax of lorenz contrast

`lorenz contrast` computes differences in Lorenz ordinates between outcome variables or subpopulations. It requires results from `lorenz estimate` to be in memory, which will be replaced by the results from `lorenz contrast`.[5] The syntax is

`lorenz contrast` [*base*] [`,` *options*]

where *base* is the name of the outcome variable or the value of the subpopulation to be used as base for the contrasts. If *base* is omitted, `lorenz contrast` computes adjacent contrasts across outcome variables or subpopulations (or contrasts with respect to the total if total results across subpopulations have been requested). Options are:

`ratio` to compute contrasts as ratios instead of differences.

`lnratio` to compute contrasts as logarithms of ratios instead of differences.

`graph`[(*options*)] to draw a line graph of the results. *options* are as described for `lorenz graph` below.

display_options such as `cformat()` or `coeflegend` to format the coefficient table. See [R] **estimation options**.

[5] Alternatively, to compute the contrasts directly, you may apply the `contrast()` option to `lorenz estimate` (see above).

3.3 Syntax of lorenz graph

`lorenz graph` draws a line diagram of Lorenz curves or Lorenz curve contrasts. It requires results from `lorenz estimate` or `lorenz contrast` to be in memory.[6] The syntax is

`lorenz graph` [`,` *options*]

where the options are as follows.

Main

`proportion` to scale the population axis as proportion (0 to 1). The default is to scale the axis as percentage (0 to 100).

`nodiagonal` to omit the equal distribution diagonal that is included by default if graphing relative Lorenz or concentration curves. No equal distribution diagonal is included if graphing equality gap curves or total, generalized, or absolute Lorenz/concentration curves or if graphing contrasts.

`diagonal(`*options*`)` to affect the rendition of the equal distribution diagonal. *options* are as described in [G] ***line_ options***.

`keep(`*list*`)` to select and order the results to be included as separate subgraphs, where *list* is a list of the names of the outcome variables or the values of the subpopulations to be included. *list* may also contain `total` for the overall results if overall results were requested. Furthermore, you may use elements such as `#1`, `#2`, `#3`, ... to refer to the 1st, 2nd, 3rd, ... outcome variable or subpopulation.

`prange(`*min max*`)` to restrict the range of the points to be included in the graph. Points whose abscissae lie outside *min* and *max* will be discarded. *min* and *max* must be within [0,100]. For example, to include only the upper half of the distribution, type `prange(50 100)`.

`gini(`*%fmt*`)` to set the format for the Gini coefficients included in the subgraph or legend labels, or `nogini` to suppresses the Gini coefficients. These options are only relevant if the `gini` option has been specified when calling `lorenz estimate`. The default format is `%9.3g`; see [R] **format**.

Labels/rendering

connect_ options to affect the rendition of the plotted lines; see [G] ***connect_ options***.

`labels("`*label 1*`" "`*label 2*`" ...)` to specify custom labels for the subgraphs of the outcome variables or subpopulations.

[6] You may also draw the graph directly by applying the `graph()` option to `lorenz estimate` or `lorenz contrast` (see above).

byopts(*byopts*) to determine how subgraphs are combined; see [G] ***by_ option***.

overlay to include results from multiple outcome variables or subpopulations in the same plot instead of creating subgraphs.

o#(*options*) to affect the rendition of the line of the #th outcome variable or subpopulation if overlay has been specified. For example, type o2(lwidth(*2)) to increase the line width for the second outcome variable or subpopulation. *options* are:

connect_ options	rendition of the plotted line (see [G] ***connect_ options***)
[no]ci	whether to draw the confidence interval
ciopts(*options*)	rendition of the confidence interval (see below)

Confidence intervals

level(#) to specify the confidence level, as a percentage, for confidence intervals. The default is the level that has been used when running lorenz estimate. level() cannot be used together with ci(bc), ci(bca), or ci(percentile).

ci(*citype*) to choose the type of confidence intervals to be plotted for results that have been computed using the bootstrap technique. *citype* may be normal (normal-based CIs; the default), bc (bias-corrected CIs) bca (bias-corrected and accelerated CIs) percentile (percentile CIs). bca is only available if BC_a confidence intervals have been requested when running lorenz estimate (see [R] **bootstrap**).

ciopts(*options*) to affect the rendition of the plotted confidence areas. *options* are as described in [G] ***area_ options***.

noci to omit confidence intervals from the plot.

Standard twoway options

addplot() to add other plots to the generated graph; see [G] ***addplot_ option***.

twoway_ options to affect the overall look of the graph, manipulate the legend, set titles, add lines, etc.; see [G] ***twoway_ options***.

4 Examples

4.1 Basic application

By default, `lorenz` computes relative Lorenz curves on a regular-grid of 20 equally spaced points across the population (plus a point at the origin). The following example shows the results for wages in the 1988 extract of the NLSW data shipped with Stata:

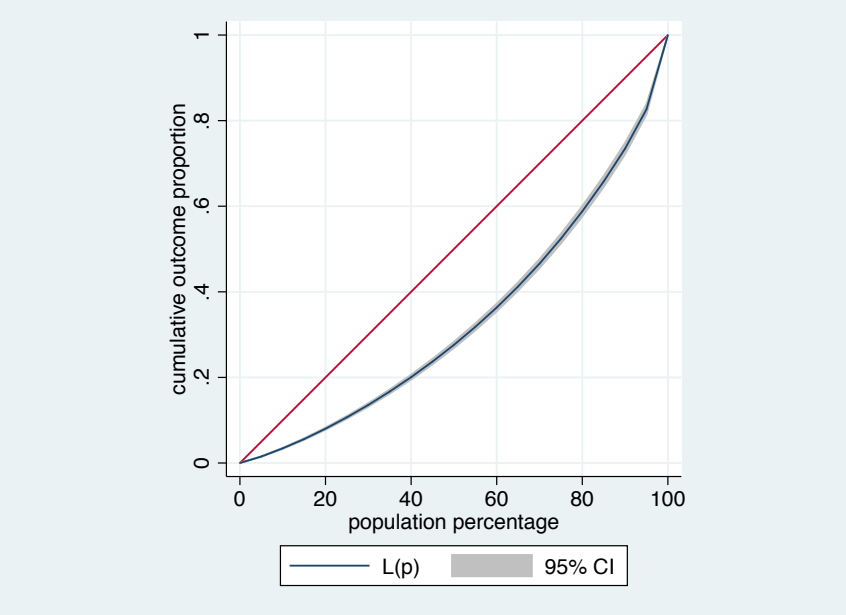
```
. sysuse nlsw88
(NLSW, 1988 extract)
. lorenz estimate wage
L(p)                                Number of obs   =      2,246
```

wage	Coef.	Std. Err.	[95% Conf.	Interval]
0	0	(omitted)		
5	.015106	.0004159	.0142904	.0159216
10	.0342651	.0007021	.0328882	.035642
15	.0558635	.0010096	.0538836	.0578434
20	.0801846	.0014032	.0774329	.0829363
25	.1067687	.0017315	.1033732	.1101642
30	.1356307	.0021301	.1314535	.1398078
35	.1670287	.0025182	.1620903	.171967
40	.2005501	.0029161	.1948315	.2062687
45	.2369209	.0033267	.2303971	.2434447
50	.2759734	.0037423	.2686347	.2833121
55	.3180215	.0041626	.3098585	.3261844
60	.3633071	.0045833	.3543191	.372295
65	.4125183	.0050056	.4027021	.4223345
70	.4657641	.0054137	.4551478	.4763804
75	.5241784	.0058003	.5128039	.5355529
80	.5880894	.0062464	.5758401	.6003388
85	.6577051	.0066148	.6447333	.6706769
90	.7346412	.0068289	.7212497	.7480328
95	.8265786	.0062687	.8142856	.8388716
100	1	.	.	.

The standard errors for the first point and the last point are zero because these Lorenz ordinates are 0 and 1 by definition. This is why Stata flags the first point as "omitted" and prints missing for the standard error and the confidence interval of the last point.

To graph the estimated lorenz curve, type:

```
. lorenz graph, aspectratio(1) xlabel(, grid)
```

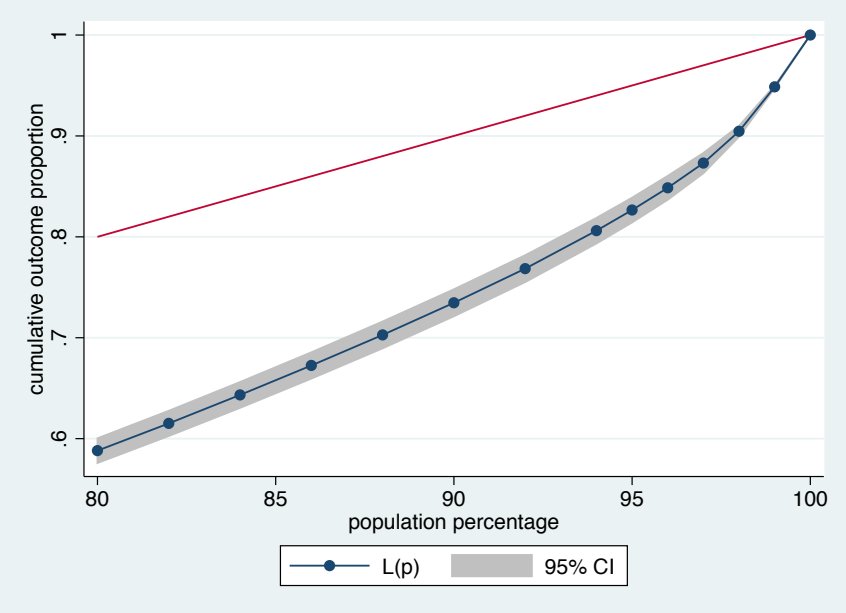


The `aspectratio(1)` option has been added to enforce a quadratic plot region and `xlabel(, grid)` has been added to include vertical grid lines. Note that, instead of applying `lorenz graph` as a separate command, you can also directly specify the `graph` option with `lorenz estimate` (see next example).

The default for `lorenz` is to use a regular grid of evaluation points across the whole distribution. To use an irregular grid or to cover just a part of the distribution, use the `percentiles()` option. The following example focusses on the upper 20% and uses a different step size towards the tail of the distribution.

```
. lorenz estimate wage, percentiles(80(2)94 95(1)100) graph(recast(connect))
L(p)                                Number of obs   =      2,246
```

wage	Coef.	Std. Err.	[95% Conf.	Interval]
80	.5880894	.0062464	.5758401	.6003388
82	.6151027	.0063755	.6026003	.6276051
84	.6432933	.0065449	.6304586	.6561281
86	.672506	.006651	.6594633	.6855487
88	.70278	.0067917	.6894613	.7160987
90	.7346412	.0068289	.7212497	.7480328
92	.7684646	.0067952	.7551391	.7817901
94	.806114	.0064727	.7934209	.8188071
95	.8265786	.0062687	.8142856	.8388716
96	.8485922	.0060386	.8367504	.860434
97	.8730971	.0051329	.8630314	.8831629
98	.9046081	.0027287	.899257	.9099591
99	.9486493	.000697	.9472826	.9500161
100	1	.	.	.



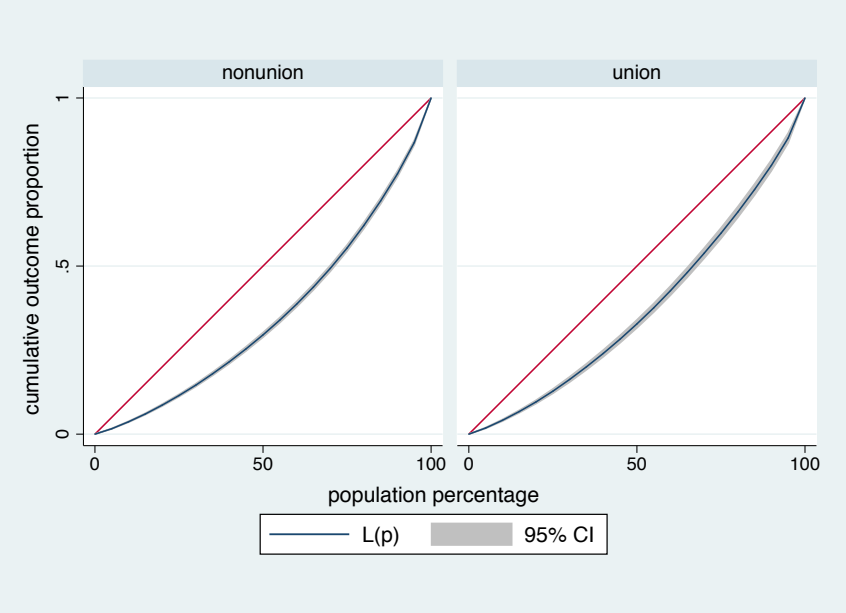
The suboption `recast(connect)` has been specified to make the evaluation points visible in the graph.

4.2 Subpopulation estimation

To compute results for multiple subpopulations, use the `over()` option. In the following example, wages are analyzed by union status:

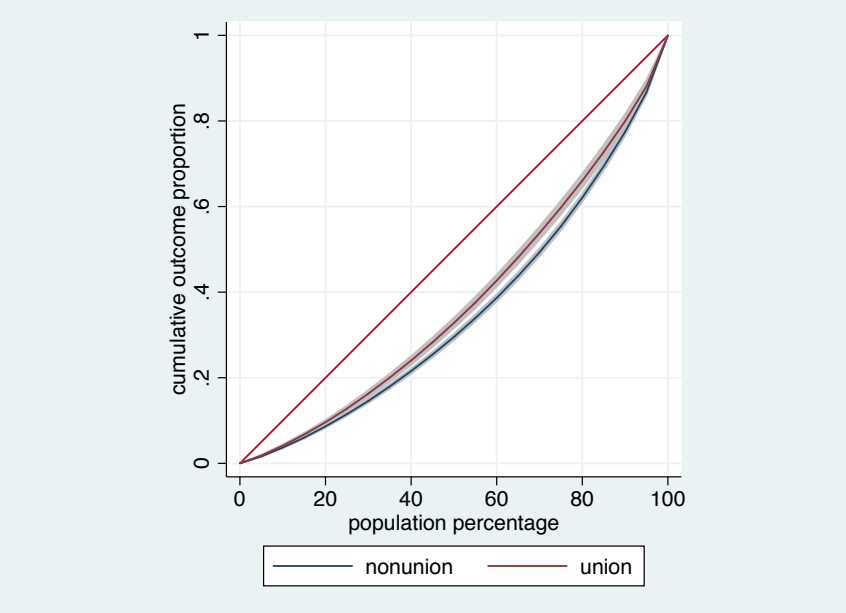
```
. lorenz estimate wage, over(union) graph(aspectratio(1))
```

(*output omitted*)



By default, `lorenz` places results for the subpopulations in separate subgraphs. To combine the results in a single plot, use the `overlay` option:

```
. lorenz graph, overlay aspectratio(1) xlabel(, grid)
```



4.3 Contrasts and Lorenz dominance

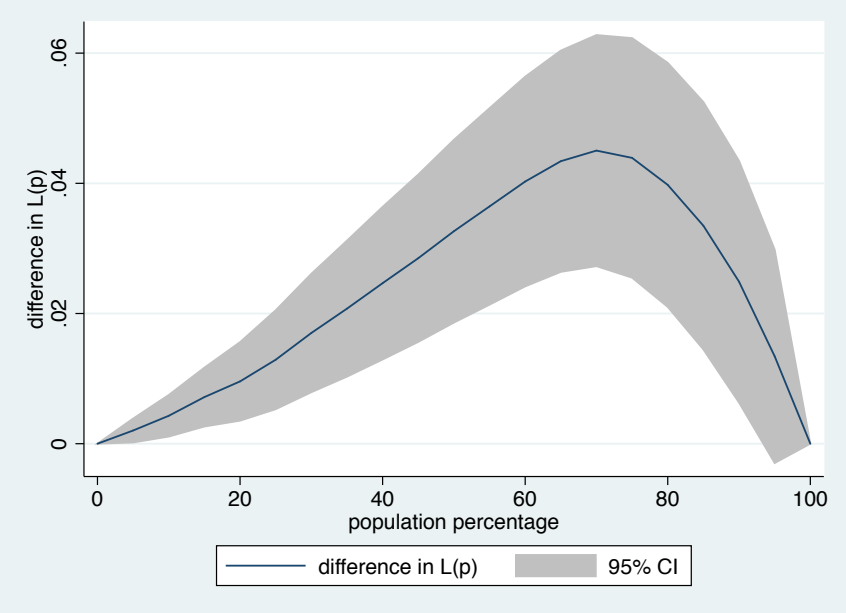
A useful feature of `lorenz` is that contrasts between subpopulations (or outcome variables) can be computed. For example, to evaluate whether the wage distribution of unionized women Lorenz dominates the wage distribution of non-unionized women, we could type:

```
. lorenz estimate wage, over(union) contrast(0) graph
L(p)                                            Number of obs     =      1,878

            0: union = nonunion
            1: union = union
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1						
0	0	(omitted)				
5	.0020273	.0009365	2.16	0.031	.0001905	.003864
10	.004292	.0016305	2.63	0.009	.0010942	.0074897
15	.0071636	.0023077	3.10	0.002	.0026376	.0116895
20	.0095728	.0030773	3.11	0.002	.0035375	.0156081
25	.0128985	.0038764	3.33	0.001	.0052959	.020501
30	.017007	.0046414	3.66	0.000	.0079042	.0261097
35	.0207488	.005331	3.89	0.000	.0102935	.031204
40	.024661	.0059814	4.12	0.000	.0129302	.0363918
45	.0284968	.0065591	4.34	0.000	.0156329	.0413607
50	.0326431	.0071647	4.56	0.000	.0185915	.0466948
55	.036453	.0077004	4.73	0.000	.0213506	.0515553
60	.0402741	.0082179	4.90	0.000	.0241569	.0563913
65	.0433946	.0086696	5.01	0.000	.0263914	.0603977
70	.0450269	.0090563	4.97	0.000	.0272654	.0627884
75	.043906	.0093882	4.68	0.000	.0254936	.0623184
80	.0397601	.009565	4.16	0.000	.021001	.0585193
85	.0334832	.0096968	3.45	0.001	.0144655	.0525008
90	.0248836	.0094742	2.63	0.009	.0063025	.0434646
95	.013423	.0083609	1.61	0.109	-.0029747	.0298208
100	0	(omitted)				

```
(difference to union = 0)
```

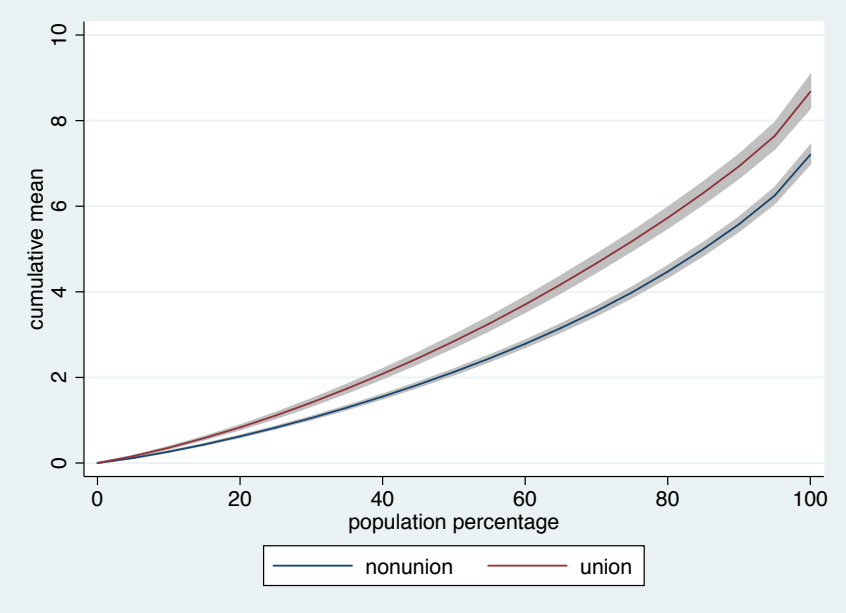


As is evident, the Lorenz curve of wages of unionized women lies above the Lorenz curve of wages of non-unionized women. Hence it appears save to conclude that the wage distribution of non-unionized women is more unequal than the wage distribution of unionized women.

Lorenz dominance does not necessarily imply that one distribution is preferable over the other from a welfare perspective. To evaluate welfare ordering, it is useful to analyze generalized Lorenz dominance. The following example shows the generalized Lorenz curves of wages of unionized and non-unionized women:

```
. lorenz estimate wage, over(union) generalized graph(overlay)
```

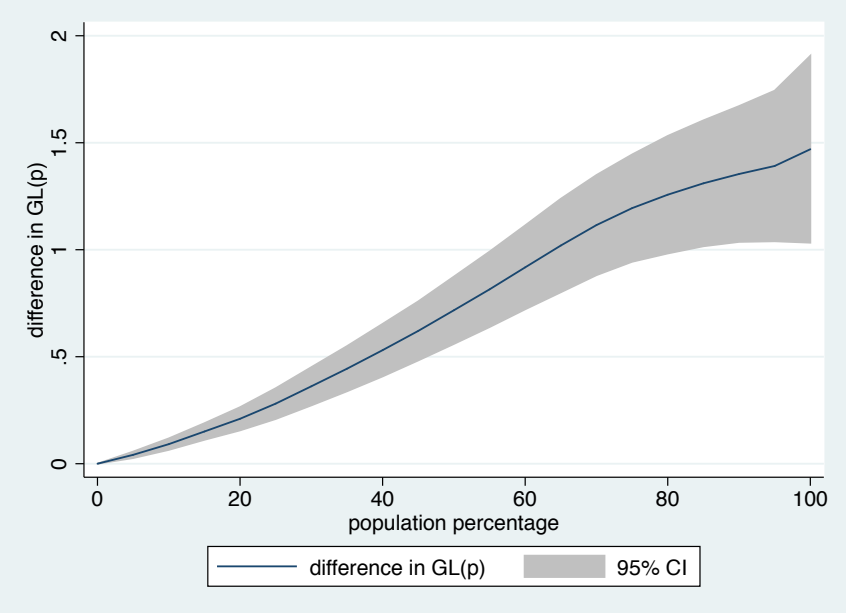
(*output omitted*)



To evaluate whether one distribution dominates the other, we can again take contrasts:

```
. lorenz contrast 0, graph
```

(*output omitted*)



It is very clear from these results, that the wage distribution of unionized women generalized Lorenz dominates the wage distribution of non-unionized women. Not only is the wage distribution of unionized women less unequal than the wage distribution of non-unionized women, it is also clearly preferable from a welfare perspective.

4.4 Concentration curves and renormalization

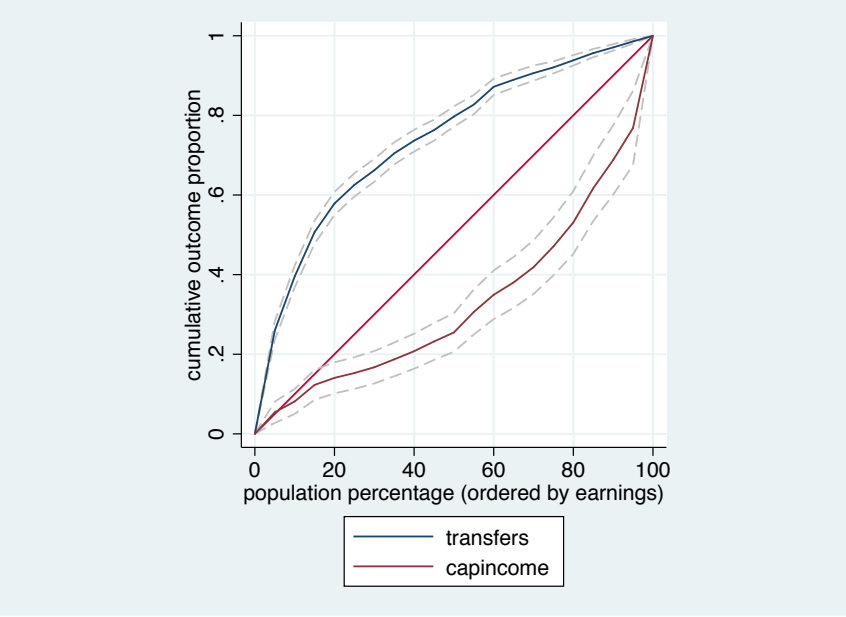
Concentration curves are used to illustrate how one variables is distributed across the population ranked by another variable. As an example, consider the following household dataset that contains information on transfer income, capital income, and earnings. We may use the `pvar()` option to analyze how transfers and capital rents are distributed across households, while ranking households by earnings:

```
. use lorenz_exampledata
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
earnings	6,007	94673.66	72285.96	0	1965925
capincome	6,007	8346.58	50226.91	0	2374227
transfers	6,007	6375.284	15838.54	0	239408

```
. lorenz transfers capincome, pvar(earnings)
>     graph(overlay aspectratio(1) xlabels(, grid) legend(cols(1))
>         ciopts(recast(rline) lp(dash)))
```

(*output omitted*)

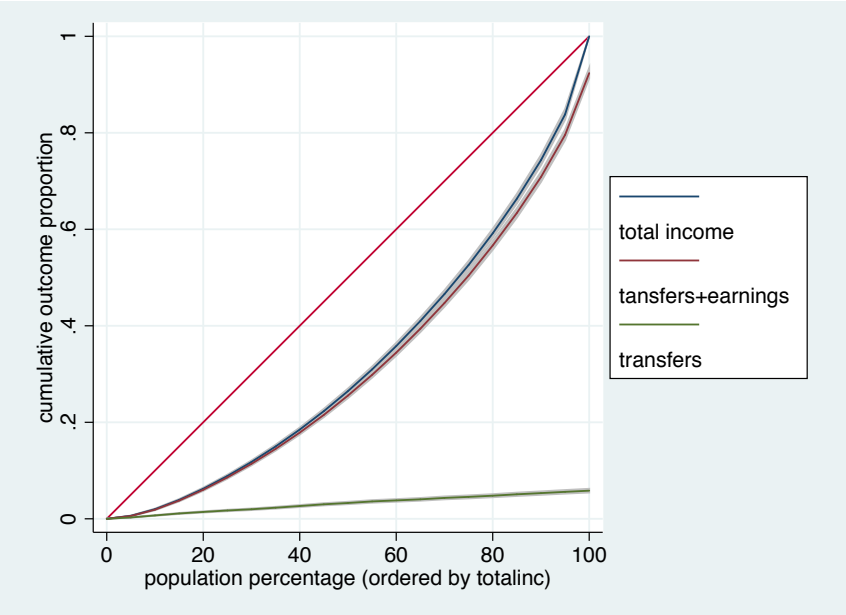


We see, as expected, that the concentration curve of transfers lies above the equal distribution line. That is, transfers benefit households with low earnings. For example, the bottom 50% of households in the earnings distribution receive about 80% of all transfers. Capital income, however, is skewed towards high earning households (the bottom 50% of households in the earnings distribution receive about 25% of all capital income).

Furthermore, the `normalize()` option can be use together with `pvar()` to subdivide the Lorenz curve of total income by income factors:

```
. generate totalinc = earnings + capincome + transfers
. generate earntrans = earnings + transfers
. lorenz totalinc earntrans transfers, pvar(totalinc) normalize(totalinc)
>     graph(overlay aspectratio(1) xlabels(, grid)
>         labels("total income" "tansfers+earnings" "transfers")
>         legend(position(3) stack cols(1)))
```

(*output omitted*)



To bottom curve displays the part of cumulative total income due to transfers. The area between the bottom curve and the middle curve depicts the contribution of earnings. The area between the middle curve and the upper curve captures the contribution of capital income.

5 Acknowledgments

This research has been supported by the Swiss National Science Foundation (Grant No. 143399).

References

Abdelkrim, A. 2005. clorenz: module to estimate Lorenz and concentration curves. Statistical Software Components S456515. Available from http://ideas.repec.org/c/boc/bocode/s456515.html.

Atkinson, A. B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Azevedo, J. P., and S. Franco. 2006. alorenz: Stata module to produce Pen's Parade, Lorenz and Generalised Lorenz curve. Statistical Software Components S456749. Available from http://ideas.repec.org/c/boc/bocode/s456749.html.

Binder, D. A., and M. S. Kovacevic. 1995. Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach. *Survey Methodology* 21(2): 137–145.

Bishop, J. A., K. V. Chow, and J. P. Formby. 1994. Testing for Marginal Changes in Income Distributions with Lorenz and Concentration Curves. *International Economic Review* 35(2): 479–488.

Cowell, F. A. 2000. Measurement of Inequality. In *Handbook of Income Distribution*, ed. A. B. Atkinson and F. Bourguignon, vol. 1, 87–166. Amsterdam: Elsevier.

———. 2011. *Measuring Inequality*. 2nd ed. Oxford: Oxford University Press.

Hao, L., and D. Q. Naiman. 2010. *Assessing Inequality*. Thousand Oaks, CA: Sage.

Hyndman, R. J., and Y. Fan. 1996. Sample Quantiles in Statistical Packages. *The American Statistician* 50(4): 361–365.

Jann, B. 2016. Assessing inequality using percentile shares. *The Stata Journal* 16(2): 264–300.

Jenkins, S. P. 2006. svylorenz: Stata module to derive distribution-free variance estimates from complex survey data, of quantile group shares of a total, cumulative quantile group shares. Statistical Software Components S456602. Available from http://ideas.repec.org/c/boc/bocode/s456602.html.

Jenkins, S. P., and P. Van Kerm. 1999. sg107: Generalized Lorenz curves and related graphs. *Stata Technical Bulletin* 48: 25–29.

Kovačević, M. S., and D. A. Binder. 1997. Variance Estimation for Measures of Income Inequality and Polarization – The Estimating Equations Approach. *Journal of Offcial Statistics* 13(1): 41–58.

Lambert, P. J. 2001. *The distribution and redistribution of income. A mathematical analysis*. 3rd ed. Manchester: Manchester University Press.

Lorenz, M. O. 1905. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association* 9(70): 209–219.

Moyes, P. 1987. A new concept of Lorenz domination. *Economics Letters* 23(2): 203–207.

Shorrocks, A. F. 1983. Ranking income distributions. *Economica* 50(197): 3–17.

Van Kerm, P., and S. P. Jenkins. 2001. Generalized Lorenz curves and related graphs: an update for Stata 7. *The Stata Journal* 1(1): 107–112.