Quantile group shares, cumulative shares (Lorenz ordinates), and generalized Lorenz ordinates: sumdist and svylorenz

Stephen P. Jenkins
ISER, University of Essex
Colchester CO4 3SQ, UK
Email: stephenj@essex.ac.uk
Overview

Extended postscript to


- Focus here on estimation of Lorenz curve and related concepts using sumdist and svylorenz

- NEW! svylorenz extended to provide variance estimates for generalized Lorenz ordinates; sumdist ported to version 8.2. (Both updated on SSC.)

- NB: recent updates on SSC also for ineqdeco, ineqdec0, povdeco, glcurve
Data for illustrations


• Available from http://www.data-archive.ac.uk/findingdata/snDescription.asp?sn=3300

• Unit record data derived from UK Family Expenditure Survey = national budget survey

• Data for 1981, 1985, 1991 used here (put in one file)
  – Income: x
  – Weight: wgt
  – Year: year
Lorenz curves and inequality

• A Lorenz curve is a plot of the cumulative income share of the poorest $100p\%$ against cumulative population share $p$, where units are ordered in ascending order of income

• Complete equality: Lorenz curve coincides with 45° ray through origin

• Inequality is greater, the further the Lorenz curve from the 45° ray

• Gini coefficient equals twice the area between the Lorenz curve and the 45° ray
Lorenz curves and inequality (2)

Axioms about inequality measures $I(x_1, x_2, \ldots, x_n)$

1. Symmetry a.k.a. Anonymity: only the income values matter, and no other information (permutation invariant)
2. Scale invariance: invariant to proportional scaling of all incomes
3. Replication Invariance: invariant to replications of the population
4. Principle of Transfers: a transfer of a small amount of income from a richer person to a poorer person (while maintaining their relative positions), reduces inequality

Lorenz dominance result (Atkinson; Foster): Lorenz curve for distribution $x$ lies on or above the Lorenz curve for $y \iff$ all inequality measures satisfying Axioms 1–4 show $I(x) < I(y)$

Derived using `glcurve` and `graph twoway`
Generalized Lorenz curves and social welfare

• Generalized Lorenz curve is the Lorenz curve scaled up at each point by population mean income, i.e. a plot of $p \mu_p$ (‘cumulative mean’) against $p$, where units are ordered in ascending order of income

• Class of social welfare functions, $\mathcal{W}_2$ with $W \in \mathcal{W}_2$ if increasing in each income, symmetric, replication-invariant and concave (i.e. a mean-preserving spread of income lowers social welfare = inequality aversion)

• Second Order Welfare Dominance result (Shorrocks): GLC($x$) above GLC($y$) at every $p \iff W(x) > W(y)$ for all $W \in \mathcal{W}_2$
  – Also implies poverty dominance by poverty gap measures
Generalized Lorenz curves (2)

\[ p \mu_p \]

Overall means shown at \( p = 1 \)

Derived using \texttt{glcurve} and \texttt{graph twoway}
Compact summaries: \texttt{sumdist}

Quantile group shares, cumulative shares (Lorenz ordinates), generalized Lorenz ordinates

\texttt{sumdist \ varname} \ [\text{aw} \ \text{fw}] \ [\text{if} \ \text{exp}] \ [\text{in} \ \text{range}] \ [, \ \_\text{n}gp(\#) \ \_\text{qgp}(\text{newvarname}) \ \_\text{p}var(\text{newvarname}) \ \_\text{l}var(\text{newvarname}) \ \_\text{gl}var(\text{newvarname})] \]

- Optional \# of quantile groups (default = 10)
- Many saved results in \texttt{r(...)}
- \texttt{by}-able
- Can derive variance estimates using \texttt{bootstrap}
- Can be used to produce variables for drawing basic Lorenz and generalized Lorenz curves (but \texttt{glcurve} is better)
**sumdist in action**

```
. sumdist x [aw= wgt] if year == 1991
```

Warning: x has 20 values = 0. Used in calculations

Distributional summary statistics, 10 quantile groups

<table>
<thead>
<tr>
<th>Quantile group</th>
<th>Quantile % of median</th>
<th>Share, %</th>
<th>L(p), %</th>
<th>GL(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92.248</td>
<td>2.961</td>
<td>2.961</td>
<td>6.923</td>
</tr>
<tr>
<td>2</td>
<td>115.768</td>
<td>4.450</td>
<td>7.411</td>
<td>17.330</td>
</tr>
<tr>
<td>3</td>
<td>141.267</td>
<td>5.469</td>
<td>12.880</td>
<td>30.121</td>
</tr>
<tr>
<td>4</td>
<td>167.221</td>
<td>6.584</td>
<td>19.465</td>
<td>45.518</td>
</tr>
<tr>
<td>5</td>
<td>194.455</td>
<td>7.732</td>
<td>27.197</td>
<td>63.600</td>
</tr>
<tr>
<td>6</td>
<td>225.385</td>
<td>9.008</td>
<td>36.204</td>
<td>84.664</td>
</tr>
<tr>
<td>7</td>
<td>263.340</td>
<td>10.407</td>
<td>46.611</td>
<td>109.001</td>
</tr>
<tr>
<td>8</td>
<td>315.397</td>
<td>12.339</td>
<td>58.950</td>
<td>137.855</td>
</tr>
<tr>
<td>9</td>
<td>402.212</td>
<td>15.145</td>
<td>74.095</td>
<td>173.272</td>
</tr>
<tr>
<td>10</td>
<td>25.905</td>
<td>100.000</td>
<td>233.852</td>
<td></td>
</tr>
</tbody>
</table>

Share = quantile group share of total x;
L(p)=cumulative group share; GL(p)=L(p)*mean(x)
Variance estimation

• Estimation using sample survey data means that estimates reflect sampling variability
• Complex survey design effects: clustering and stratification also affect sampling variability
• Relatively neglected topic in income distribution analysis to date:
  – Non-sampling issues viewed as mattering more?
  – Large samples argument about SEs likely to be small
    • But what about subgroups? What is ‘large’?
  – Appropriate software previously unavailable … but is now for many of the methods used
    • Focus on linearization methods here
Variance estimation methods

- Beach and Davidson (1983): formulae for variance estimation of shares, cumulative shares and generalized Lorenz ordinates, but for unweighted data with no complex survey design features
- Beach and Kaliski (1986): extend results to the case with sample weights that are fixed and non-stochastic
- \texttt{svylorenz} variance estimates:
  - Cumulative shares and Gini: Kovacevic and Binder (1997)
  - Quantile group shares: Beach and Kaliski (1986) result relating variances shares to variances of cumulative shares
Estimation using *svylorenz*

Quantile group shares, cumulative shares (Lorenz ordinates), generalized Lorenz ordinates, and Gini coefficient

```
svylorenz varname [if exp] [in range] [, ngp(#) qgp(newvarname)
subpop(varname) pvar(newvarname)
lvar(newvarname) selvar(newvarname)
glvvar(newvarname)
seglvvar(newvarname) level(#) ]
```

- Data must be *svyset* before using this command
- Optional # of quantile groups (default = 10)
- Many saved results in e(…)

*University of Essex*
Assumptions about survey design in the ‘IFS’ dataset

• There are no PSU or strata variables supplied in the IFS data

• However, the observations (families) are clustered in households (= sampling unit):
  - each person in each family is assumed to have the income of household to which s/he belongs

• Estimate variances accounting for within-household clustering, and the weighting
  - `svyset hrn [pw = wgt]`
Estimation using `svylorenz`

`. svylorenz x if year == 1991`

Warning: x has 20 values = 0. Used in calculations

Quantile group shares, cumulative shares (Lorenz ordinates),
generalized Lorenz ordinates, and Gini

Number of strata = 1
Number of PSUs = 9772
Number of obs = 9772
Population size = 54872650.00
Design df = 9771

| Group | Linearized share | Estimate | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|-------|------------------|----------|-----------|-------|-----|-----------------------|
| 1     | 0.029606         | 0.010052 | 2.945     | 0.003 | .0099039 | .0493083 |
| 2     | 0.044503         | 0.000596 | 74.629    | 0.000 | .0433338 | .0456714 |
| 3     | 0.054694         | 0.000793 | 68.952    | 0.000 | .0531389 | .0562483 |
| 4     | 0.065844         | 0.000908 | 72.522    | 0.000 | .0640648 | .0676238 |
| 5     | 0.077321         | 0.001303 | 79.876    | 0.000 | .0753555 | .0792859 |
| 6     | 0.090076         | 0.001136 | 79.280    | 0.000 | .0878488 | .0923025 |
| 7     | 0.104067         | 0.001303 | 79.876    | 0.000 | .101513  | .106622  |
| 8     | 0.123386         | 0.001566 | 78.777    | 0.000 | .120316  | .126456  |
| 9     | 0.151451         | 0.002019 | 75.012    | 0.000 | .147494  | .155408  |
| 10    | 0.259053         | 0.006431 | 40.283    | 0.000 | .246449  | .271657  |

Default number of quantile groups = 10;
number can be chosen by the user
## Variance estimation (continued)

<table>
<thead>
<tr>
<th>Cumul. share</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.029606</td>
<td>0.010052</td>
<td>2.945</td>
<td>0.003</td>
<td>.0099039</td>
<td>.0493083</td>
</tr>
<tr>
<td>2</td>
<td>0.074109</td>
<td>0.009867</td>
<td>7.511</td>
<td>0.000</td>
<td>.0547693</td>
<td>.0934482</td>
</tr>
<tr>
<td>3</td>
<td>0.128802</td>
<td>0.009594</td>
<td>13.425</td>
<td>0.000</td>
<td>.109999</td>
<td>.147606</td>
</tr>
<tr>
<td>4</td>
<td>0.194647</td>
<td>0.009265</td>
<td>21.010</td>
<td>0.000</td>
<td>.176488</td>
<td>.212805</td>
</tr>
<tr>
<td>5</td>
<td>0.271967</td>
<td>0.008885</td>
<td>30.609</td>
<td>0.000</td>
<td>.254553</td>
<td>.289382</td>
</tr>
<tr>
<td>6</td>
<td>0.362043</td>
<td>0.008445</td>
<td>42.871</td>
<td>0.000</td>
<td>.345491</td>
<td>.378595</td>
</tr>
<tr>
<td>7</td>
<td>0.466110</td>
<td>0.007917</td>
<td>58.875</td>
<td>0.000</td>
<td>.450593</td>
<td>.481627</td>
</tr>
<tr>
<td>8</td>
<td>0.589496</td>
<td>0.007275</td>
<td>81.035</td>
<td>0.000</td>
<td>.575238</td>
<td>.603754</td>
</tr>
<tr>
<td>9</td>
<td>0.740947</td>
<td>0.006431</td>
<td>115.219</td>
<td>0.000</td>
<td>.728343</td>
<td>.753551</td>
</tr>
<tr>
<td>10</td>
<td>1.000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gen. Lorenz</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.923</td>
<td>0.158</td>
<td>43.861</td>
<td>0.000</td>
<td>6.614</td>
<td>7.233</td>
</tr>
<tr>
<td>2</td>
<td>17.330</td>
<td>0.249</td>
<td>69.703</td>
<td>0.000</td>
<td>16.843</td>
<td>17.818</td>
</tr>
<tr>
<td>3</td>
<td>30.121</td>
<td>0.377</td>
<td>79.888</td>
<td>0.000</td>
<td>29.382</td>
<td>30.860</td>
</tr>
<tr>
<td>4</td>
<td>45.518</td>
<td>0.524</td>
<td>86.828</td>
<td>0.000</td>
<td>44.491</td>
<td>46.546</td>
</tr>
<tr>
<td>5</td>
<td>63.600</td>
<td>0.683</td>
<td>93.104</td>
<td>0.000</td>
<td>62.261</td>
<td>64.939</td>
</tr>
<tr>
<td>6</td>
<td>84.664</td>
<td>0.860</td>
<td>98.495</td>
<td>0.000</td>
<td>82.980</td>
<td>86.349</td>
</tr>
<tr>
<td>7</td>
<td>109.001</td>
<td>1.057</td>
<td>103.169</td>
<td>0.000</td>
<td>106.930</td>
<td>111.071</td>
</tr>
<tr>
<td>8</td>
<td>137.855</td>
<td>1.290</td>
<td>106.900</td>
<td>0.000</td>
<td>135.327</td>
<td>140.382</td>
</tr>
<tr>
<td>9</td>
<td>173.272</td>
<td>1.576</td>
<td>109.926</td>
<td>0.000</td>
<td>170.182</td>
<td>176.361</td>
</tr>
<tr>
<td>10</td>
<td>233.852</td>
<td>2.711</td>
<td>86.247</td>
<td>0.000</td>
<td>228.538</td>
<td>239.166</td>
</tr>
</tbody>
</table>

Gini estimates are based on the complete unit record data (not grouped data)
Lorenz curve comparisons with CIs

```
. svylorenz x if year == 1981, pvar(p81) lvar(rl81) selvar(se81)
. svylorenz x if year == 1991, pvar(p91) lvar(rl91) selvar(se91)

. local half_alpha = (1 - `c(level)' / 100) / 2
. gen lcl81 = rl81  + invnorm(`half_alpha') * se81
  (25222 missing values generated)
. gen ucl81 = rl81  + invnorm(1-`half_alpha') * se81
  (25222 missing values generated)
. gen lcl91 = rl91  + invnorm(`half_alpha') * se91
  (25222 missing values generated)
. gen ucl91 = rl91  + invnorm(1-`half_alpha') * se91
  (25222 missing values generated)

. graph twoway (connect rl81 p81, sort yaxis(1 2) )                       ///
   (connect rl91 p91, sort yaxis(1 2) )                    ///
   (function y = x, range(0 1) yaxis(1 2) )                ///
   (rspike lcl81 ucl81 p81, blcolor(black) sort ) ///
   (rspike lcl91 ucl91 p91, blcolor(black) sort ) ///
   , aspect(1) xtitle("Cumulative population share, p")    ///
   ytitle("Cumulative income share, poorest 100p%", axis(1)) ytitle(" ",
   axis(2)) ///
   legend(label (1 "1981") label(2 "1991") label(3 "Equality") ///
   label(4 "95%CI,1981") label(5 "95%CI,1991") size(small) ///
   region(lstyle(none)) ) saving(svylorenz81_91, replace)
  (file svylorenz81_91.gph saved)
```

NB graphs can also be derived using glcurve, but no CIs
Lorenz curve comparisons with CIs (2)

Note overlapping CIs at small values of $p$
Further issues

• multiple comparison tests given a set of (generalized) Lorenz estimates
  – stochastic dominance checks
  – See discussion and references in e.g. Davidson and Duclos (2000)
Bootstrap methods

A general empirically-based approach which you may prefer, because:

- Linearization method may be too complicated for your application, and/or software unavailable
- All the linearization sampling variance formulae are ‘approximate’, large sample, formulae and you may not trust them
- It is very flexible in principle
  - But is no panacea: requires careful set-up for complex survey designs other than those that bootstrap options allow
Bootstrapped SE for Gini index

1. Write wrapper program to retrieve results from `ineqdec0`
   - `svylorenz` uses obs with values $\geq 0$, `ineqdeco` uses obs with values $> 0$, `ineqdec0` and `sumdist` uses obs with any real value

2. Drop observations not to be used in the bootstrapping
   - Apply similar methods to derive bootstrap estimates from any other program producing estimates of inequality measures (including Lorenz ordinates)

```
. cap prog drop ineq

. prog define ineq, rclass
   1.     ineqdec0 x [aw = wgt]
   2.     ret scalar gini = r(gini)
   3.     end

. drop if (missing(x) | x < 0 | year != 1991 )
(18783 observations deleted)
```
Bootstrapped SEs for inequality indices (2)

. * 250 reps
. bootstrap gini = r(gini), reps(250) cluster(hrn) : ineq
(running ineq on estimation sample)
<output omitted>

Bootstrap results

| Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|-------|-----------|------|-----|----------------------|
| gini  | .3365993  | .0045669 | 73.70 | 0.000 | .3276483  .3455502 |

Bootstrap SE is similar to the linearized estimate from svylorenz:

| Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|-------|-----------|------|-----|----------------------|
| Gini  | 0.3365993 | .00515134 | 65.342 | 0.000 | .3265028  .3466957 |
Bootstrapped SEs for shares, etc.

Use similar estimation strategy:

```
. cap prog drop sdist

. prog define sdist, rclass
  1.    sumdist x [aw = wgt]
  2.    ret scalar sh10 = r(sh10)
  3.    end

. drop if (missing(x) | x < 0 | year != 1991)
(18783 observations deleted)
```
Bootstrapped SEs for shares, etc. (2)

. * 250 reps
. bootstrap sharetop10pc = r(sh10), reps(250) cluster(hrn) : sdist
(running sdist on estimation sample)

<output omitted>

Bootstrap results                               Number of obs =     25231
Number of clusters =     19702
Replications    =       250

command:  sdist
sharetop10pc:  r(sh10)

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Bootstrap</th>
<th>Normal-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>z</td>
</tr>
<tr>
<td>sharetop10pc</td>
<td>.2590531</td>
<td>.0054408</td>
<td>47.61</td>
</tr>
</tbody>
</table>

Bootstrap SE is similar to the linearized estimate from svylorenz:

|                | Coef.    | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|----------------|----------|-----------|-----|-----|---------------------|
| sharetop10pc   | 0.259053 | 0.006431  | 40.283 | 0.000 | 0.246449 .271657 |
Advertisement
Suite of Stata programs for analysis of distributions, also with variance estimation

- Available from SSC or as Stata Journal update
  - ineqdeco, ineqdec0, povdeco
    - Variance estimates via the bootstrap
  - svyatk, svygei
    - Variance estimates via linearization
  - glcurve (joint with Philippe Van Kerm)
    - Draw (generalized) Lorenz, concentration, TIP curves, etc.
  - svylorenz (and sumdist)
    - Variance estimates via linearization (and bootstrap)

NB All rely on you having ‘good’ data and making appropriate choices about definitions of ‘income’, the income-receiving ‘unit’, etc.!
References


