University of Essex

UKSUG 2006

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

ISER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

Overview

Extended postscript to

- Jenkins, S.P. 2006. Estimation and interpretation of measures of inequality, poverty, and social welfare using Stata. Presentation at North American Stata Users' Group Meetings 2006, Boston MA. http://econpapers.repec.org/paper/bocasug06/16.htm
- 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

"Institute for Fiscal Studies (IFS) 'Households Below Average Income Dataset', 1961-1991" data

- 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

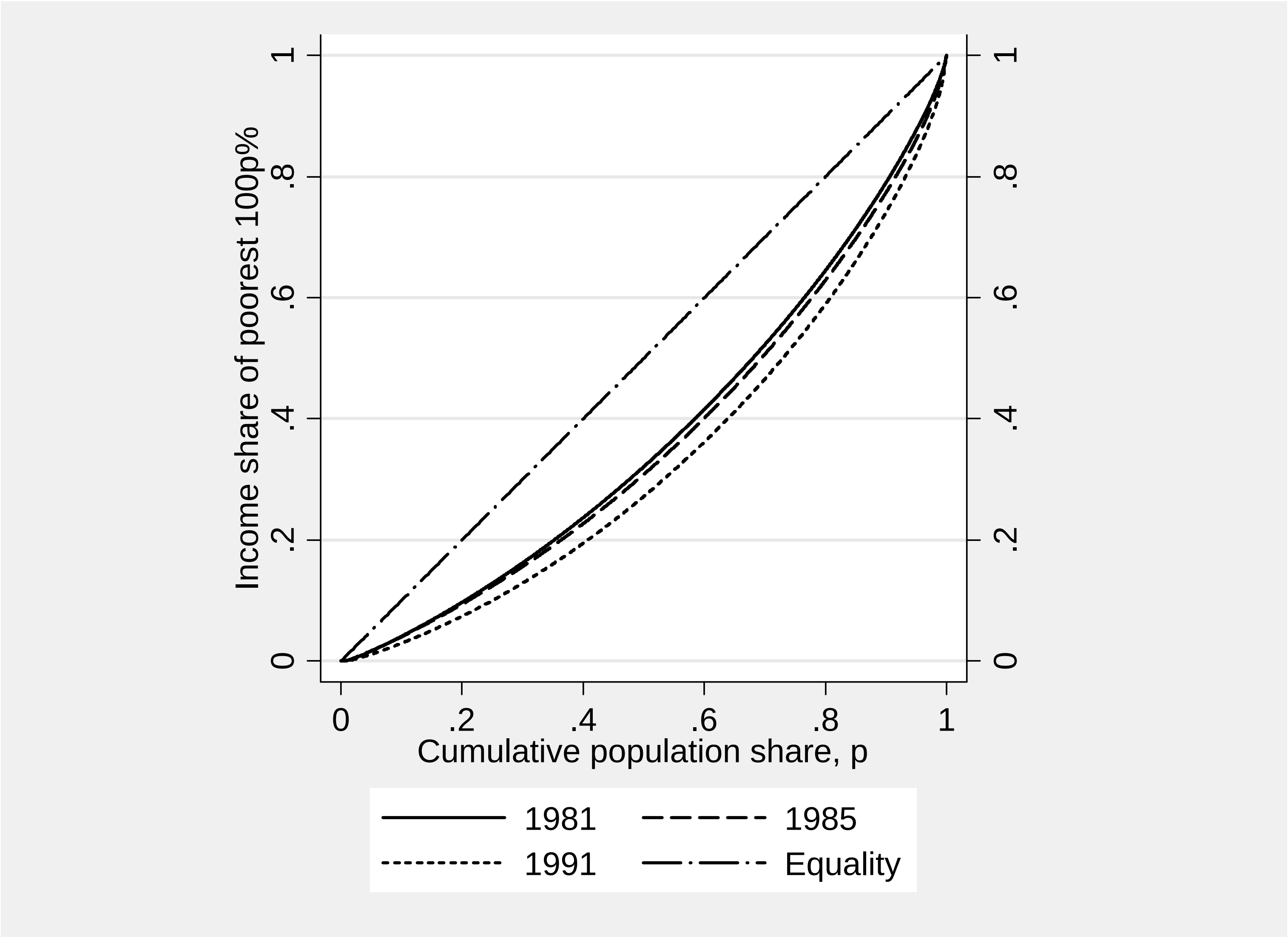
Lorenz curves and inequality (2)

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 $\boldsymbol{x}$ lies on or above the Lorenz curve for $\boldsymbol{y}$ $\Leftrightarrow$ all inequality measures satisfying Axioms 1–4 show $I(\boldsymbol{x}) < I(\boldsymbol{y})$

Inequality comparisons: 1981, 1985, 1991



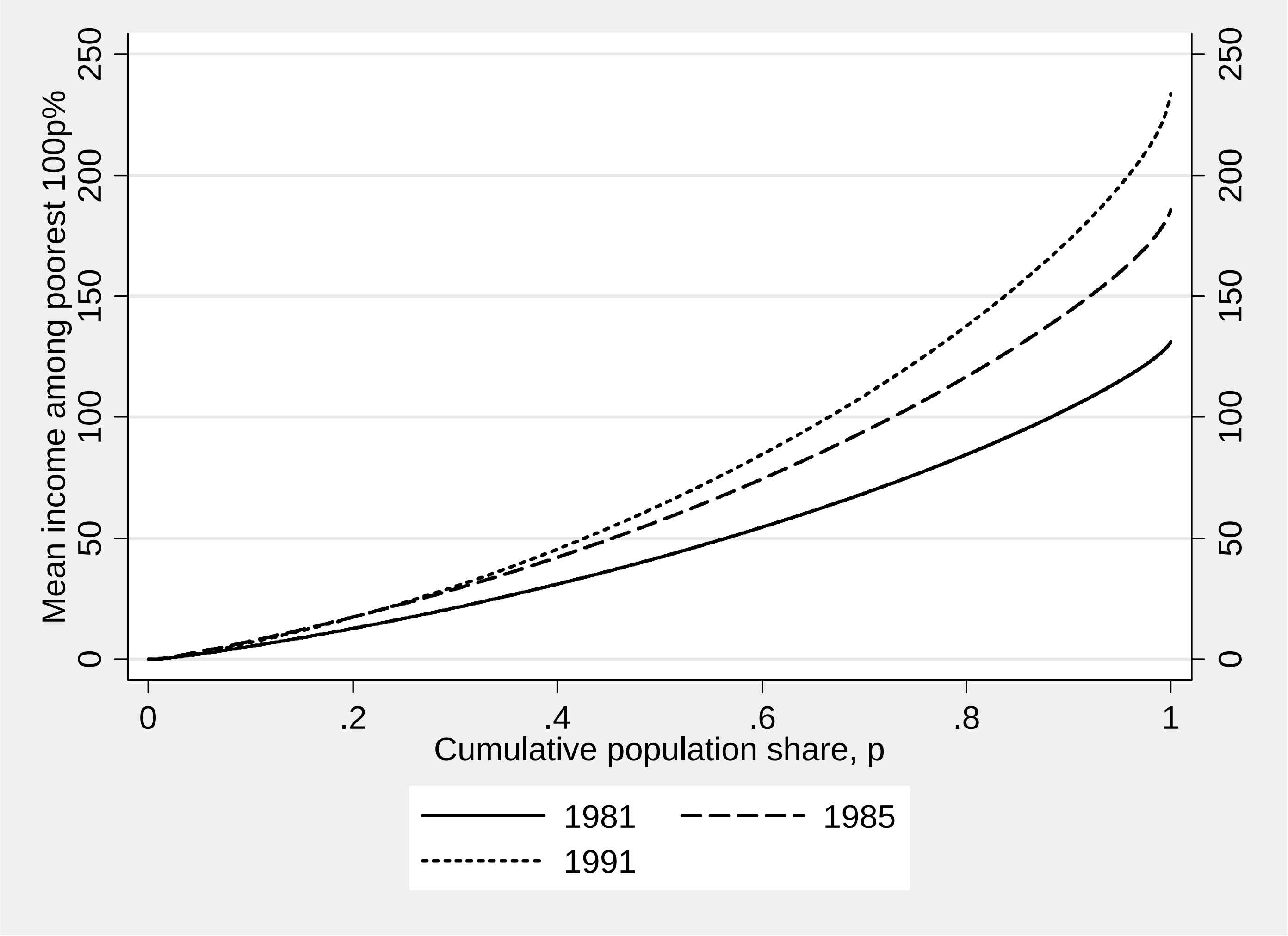
Derived using `glcurve` and `graph twoway`

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($\boldsymbol{x}$) above GLC($\boldsymbol{y}$) at every $p \Leftrightarrow W(\boldsymbol{x}) > W(\boldsymbol{y})$ 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 `glcurve` and `graph twoway`

Compact summaries: `sumdist`

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

`sumdist` *varname* `[aw fw] [if` *exp*`] [in` *range*`] [, ngp(#) qgp(`*newvarname*`) pvar(`*newvarname*`) lvar(`*newvarname*`) glvar(`*newvarname*`)]`

- Optional # of quantile groups (default = 10)
- Many saved results in `r(`...`)`
- `by`-able
- Can derive variance estimates using `bootstrap`
- Can be used to produce variables for drawing basic Lorenz and generalized Lorenz curves (but `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

---------------------------------------------------------------------------
Quantile  |
group     |   Quantile  % of median     Share, %      L(p), %        GL(p)
----------+----------------------------------------------------------------
        1 |     92.248       47.439        2.961        2.961        6.923
        2 |    115.768       59.534        4.450        7.411       17.330
        3 |    141.267       72.648        5.469       12.880       30.121
        4 |    167.221       85.995        6.584       19.465       45.518
        5 |    194.455      100.000        7.732       27.197       63.600
        6 |    225.385      115.906        9.008       36.204       84.664
        7 |    263.340      135.424       10.407       46.611      109.001
        8 |    315.397      162.195       12.339       58.950      137.855
        9 |    402.212      206.841       15.145       74.095      173.272
       10 |                               25.905      100.000      233.852
---------------------------------------------------------------------------
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
- Binder & Kovacevic (1995) and Kovacevic & Binder (1997): 'estimating equations' approach yields formulae for variance estimation of cumulative shares and shares, and Gini coefficient, allowing for probability weights and for complex survey design more generally. See also Zheng (2002) .
- `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
  - Generalized Lorenz ordinates: SPJ's application of the estimating equations approach of Binder and Kovacevic (1995) and Kovacevic and Binder (1997)

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)
  glvar(newvarname)
  seglvar(newvarname) level(#) ]
```

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

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`

Default number of quantile groups = 10; number can be chosen by the user

```
. 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 obs     =        9772
Number of PSUs   =       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   0.001003   77.115    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      .10662
     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
----------+---------------------------------------------------------------
```

Variance estimation (continued)

Cumul. share						
1	0.029606	0.010052	2.945	0.003	.0099039	.0493083
2	0.074109	0.009867	7.511	0.000	.0547693	.0934482
3	0.128802	0.009594	13.425	0.000	.109999	.147606
4	0.194647	0.009265	21.010	0.000	.176488	.212805
5	0.271967	0.008885	30.609	0.000	.254553	.289382
6	0.362043	0.008445	42.871	0.000	.345491	.378595
7	0.466110	0.007917	58.875	0.000	.450593	.481627
8	0.589496	0.007275	81.035	0.000	.575238	.603754
9	0.740947	0.006431	115.219	0.000	.728343	.753551
10	1.000000					
Gen. Lorenz						
1	6.923	0.158	43.861	0.000	6.614	7.233
2	17.330	0.249	69.703	0.000	16.843	17.818
3	30.121	0.377	79.888	0.000	29.382	30.860
4	45.518	0.524	86.828	0.000	44.491	46.546
5	63.600	0.683	93.104	0.000	62.261	64.939
6	84.664	0.860	98.495	0.000	82.980	86.349
7	109.001	1.057	103.169	0.000	106.930	111.071
8	137.855	1.290	106.900	0.000	135.327	140.382
9	173.272	1.576	109.926	0.000	170.182	176.361
10	233.852	2.711	86.247	0.000	228.538	239.166
Gini	0.3365993	.00515134	65.342	0.000	.3265028	.3466957

Gini estimates are based on the complete unit record data (not grouped data)

Lorenz curve comparisons with CIs

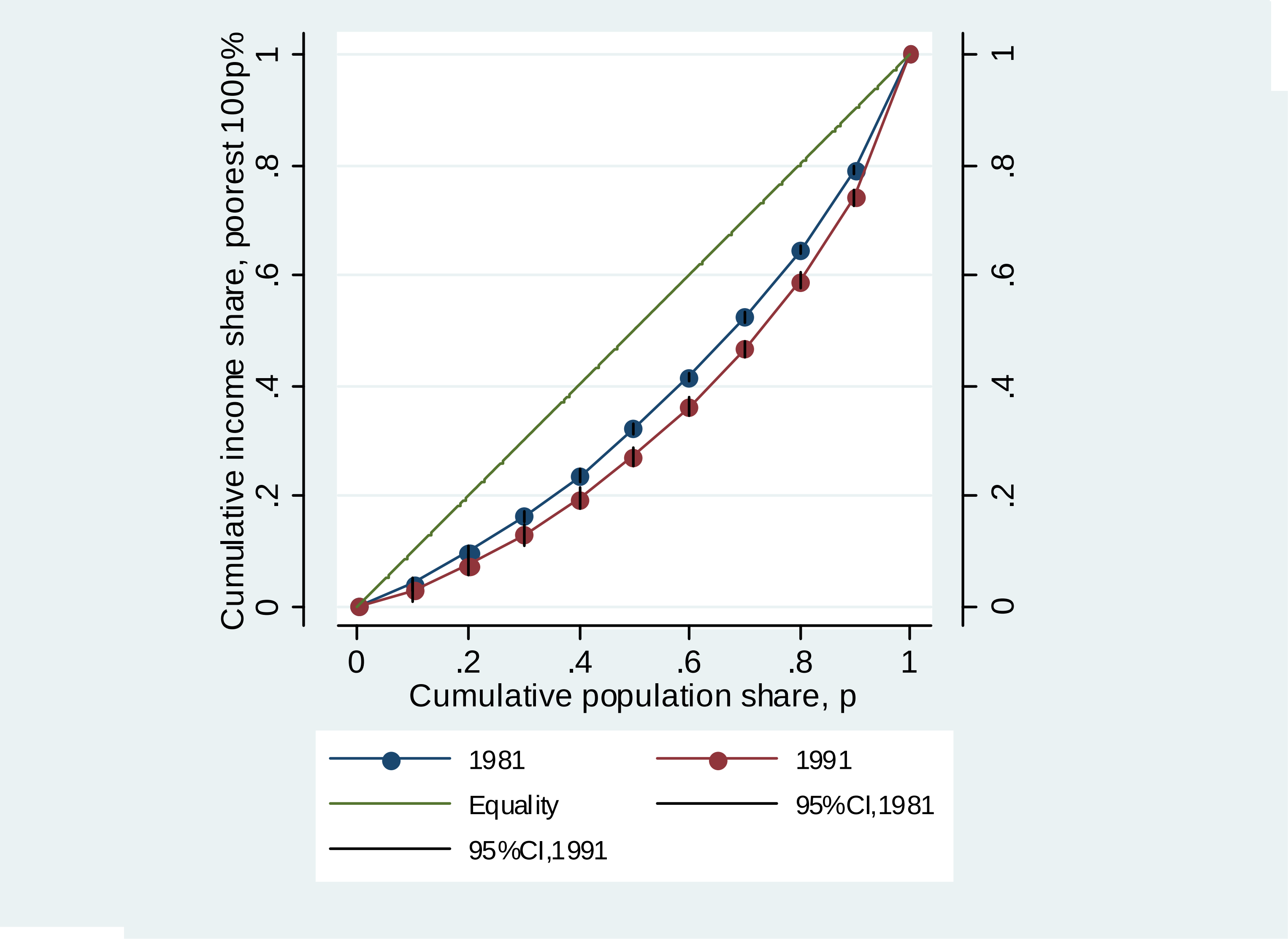
```
. svylorenz x if year == 1981, pvar(p81) lvar(rl81) selvar(se81)
. svylorenz x if year == 1991, pvar(p91) lvar(rl91) selvar(se91)

. 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 ≥ 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                               Number of obs      =      6468
                                                Number of clusters =      5254
                                                Replications       =       250

      command:  ineq
         gini:  r(gini)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      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`:

```
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)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sharetop10pc |   .2590531   .0054408    47.61   0.000     .2483894    .2697168
------------------------------------------------------------------------------
```

Bootstrap SE is similar to the linearized estimate from `svylorenz`:

```
             |   0.259053   0.006431   40.283   0.000     .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

Beach, C.M. and R. Davidson. 1983. Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies* 50: 723–725.

Beach, C.M. and S.F. Kaliski. 1986. Lorenz curve inference with sample weights: an application to the distribution of unemployment experience. *Applied Statistics* 35(1): 38–45.

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: 137-145.

Davidson, R. and J-Y. Duclos. 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* 68: 1435–1464

Kovaevic, M.S. and D.A. Binder. 1997. Variance estimation for measures of income inequality and polarization. *Journal of Official Statistics* 13(1): 41–58. Full text downloadable: http://www.jos.nu/Articles/abstract.asp?article=1314

Zheng, B. 2002. Testing Lorenz curves with non-simple random samples. *Econometrica* 70: 1235–1243.