NASUG 2006

Estimation and interpretation of measures of inequality, poverty and social welfare using Stata

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

The focus and perspective

- Focus on *income* distributions, but methods applicable to many other variables
- Perspective of an *economist* (and one particular economist at that), but informed by other disciplines
- Focus on cross-sectional perspectives rather than income dynamics
- 'Descriptive' analysis; no multivariate modelling
- Emphasis on applications rather than theoretical detail
- Uses my programs; there are others
- Multiple perspectives on and answers to the question: "How did the UK income distribution change between 1981 and 1991?"
- Log files will be made available on the Meetings webpage

What’s wrong with using the variance for distributional comparisons?

- Nothing really, except that …
- It summarizes dispersion in a particular way and we may not like the properties it has
  - E.g. unlike the CV, the variance is not invariant to an equiproportionate change in each value (cf. effects of price inflation on money incomes when studying trends over time in inequality)
- We may be interested in other distributional features besides inequality, e.g. poverty and social welfare

Methodological approach

- Seek to make comparisons robust to differences in views about how one should summarize the various 'economic' features of the distributions

$\Rightarrow$

- *Dominance checks*: methods for deriving conclusions about distributional comparisons that are robust to differences in views about e.g. how averse you are to inequality, poverty, etc.
- *Summary indices* incorporating different views parametrically

Outline

- Summarizing and picturing distributions
  - Getting to know your data
  - Density estimation and subgroup decomposition
  - Pen’s Parade
  - Lorenz and generalized Lorenz curves
  - Three ‘I’s of Poverty curves
- Summary indices of inequality, poverty, social welfare, with decompositions by subgroup
- Variance estimation to account for sampling variability

Application of any statistical methods is predicated on a number of important choices

Checklist	***Choices used here***
• Reference unit for income	Household
• Observation unit in data	Nuclear family ('benefit unit')
• Equivalence scale	UK 'McClements BHC' scale
• Weighting of observations	Distribution among individuals
• Concept of resources	Net (= post-tax post-transfer) income
• Time period	'Current' (rather than annual)
• Price deflator(s)	Convert to 1991 prices using RPI
• Absolute or relative?	Relative income differences
• Poverty line	60% contemporary median income

- The choices reflect commonly-used European conventions and data availability
- Assessment of their validity depends on social judgements, not only statistical issues
- Different choices from the Checklist can have large and systematic effects on results!

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`

Preliminary checks and summaries using built-in commands

- Missing values
  - None in this data set (imputation flags not included!)
- High- and low-income outliers, including
- Zero and negative income values

```
sort year
by year: count if missing(x)

by year: count if x < 0

by year: count if x == 0

by year: count if x > 0 & x < 1
```

1981	1985	1991
0	0	0
0	0	0
20	20	20
4	2	1

Summary statistics, 1981

```
. summarize x [aw=wgt] if year == 1981, detail

                    Equiv. net income, £ p.w.
-------------------------------------------------------------
      Percentiles      Smallest
 1%     29.47834              0
 5%     59.45737              0
10%     67.35584              0       Obs                9772
25%     85.71757              0       Sum of Wgt.    54872650

50%     117.1399                      Mean           131.2275
                        Largest       Std. Dev.      66.92031
75%     160.4577       627.5959
90%     211.5341       627.5959       Variance       4478.328
95%     246.7837       925.5372       Skewness        2.05027
99%     369.8264        931.272       Kurtosis       11.68375
```

Summary statistics, 1991

```
. summarize x [aw=wgt] if year == 1991, detail

                    Equiv. net income, £ p.w.
-------------------------------------------------------------
      Percentiles      Smallest
 1%        29.04              0
 5%     78.43056              0
10%     92.24827              0       Obs                6468
25%     127.3074              0       Sum of Wgt.    55851705

50%     194.4551                      Mean           233.8519
                        Largest       Std. Dev.      173.9472
75%     287.2739       1667.386
90%      402.212       1667.871       Variance       30257.64
95%     503.1029       1671.879       Skewness       3.677977
99%     942.0244       2734.264       Kurtosis       26.84612
```

Some basic summary statistics: CV = SD/mean, percentile ratio *p*90/*p*10

```
. qui summarize x if year == 1991, detail
. local cv_91 = r(sd)/r(mean)
. local r9010_91 = r(p90)/r(p10)
. local z_91 = 0.6 * r(p50)   // poverty line (see below)
```

And similarly for 1985 and 1981

Calculations using trimmed data may be informative about the impact of high and low income outliers (the issue of whether to always trim is not considered here!)

```
* trimming top 1% and bottom 1% of observations
. qui summarize x [aw=wgt] if year == 1991, detail
. summarize x [aw=wgt] if x > r(p1) & x < r(p99) & year == 1991,
  de
```

And similarly for 1985 and 1981

Raw versus trimmed summary statistics

RAW	1981	1985	1991
Mean	131.2	185.7	233.9
Median	117.1	161.3	194.5
CV	0.530	0.560	0.714
*p*90/*p*10	3.191	3.240	4.329

TRIM MED	1981	1985	1991
Mean	129.1	181.2	224.9
Median	117.1	161.3	194.5
CV	0.439	0.481	0.580
*p*90/*p*10	3.029	3.151	4.162

NB I use non-trimmed distributions from here onwards

Proportion poor (poverty line = 60% contemporary median income)

```
. display as text "Poverty line 1981 =  " as result `z_81'
Poverty line 1981 =  72.34352
. display as text "Poverty line 1985 =  " as result `z_85'
Poverty line 1985 =  98.417899
. display as text "Poverty line 1991 =  " as result `z_91'
Poverty line 1991 =  116.22643
. * poverty status (income below 60% of contemporary median)
. gen poor = (year==1981)*(x < `z_81' ) + (year==1985)*(x <
   `z_85' ) +  (year==1991)*(x < `z_91' ) if x < .
. lab var poor "Income < 60% median“
.  tab year poor [aw=wgt], row nofreq
    survey |  Income < 60% median
      year |         0          1 |     Total
-----------+----------------------+----------
      1981 |     85.90      14.10 |    100.00
      1985 |     86.24      13.76 |    100.00
      1991 |     79.79      20.21 |    100.00
-----------+----------------------+----------
     Total |     83.95      16.05 |    100.00
```

Income shares etc.: sumdist

```
. sumdist x [aw= wgt] if year == 1981, ng(5)

Warning: x has 20 values = 0. Used in calculations
Distributional summary statistics, 5 quantile groups

---------------------------------------------------------------------------
Quantile  |
group     |     Quantile  % of median     Share, %      L(p), %        GL(p)
----------+----------------------------------------------------------------
        1 |        79.66        68.00         9.71         9.71        12.75
        2 |       104.08        88.85        13.97        23.69        31.08
        3 |       131.62       112.36        17.91        41.59        54.58
        4 |       172.74       147.46        22.91        64.51        84.65
        5 |                                 35.49       100.00       131.23
---------------------------------------------------------------------------
Share = quantile group share of total x;
L(p)=cumulative group share; GL(p)=L(p)*mean(x)
```

`sumdist` has options for choice of the number of quantile groups used (default = 10), and to create quantile group membership variable

… and again for 1991

```
. sumdist x [aw= wgt] if year == 1991, ng(5)

Warning: x has 20 values = 0. Used in calculations
Distributional summary statistics, 5 quantile groups

---------------------------------------------------------------------------
Quantile  |
group     |    Quantile  % of median     Share, %      L(p), %       GL(p)
----------+----------------------------------------------------------------
        1 |      115.77        59.53         7.41         7.41       17.33
        2 |      167.22        85.99        12.05        19.46       45.52
        3 |      225.39       115.91        16.74        36.20       84.66
        4 |      315.40       162.20        22.75        58.95      137.85
        5 |                                 41.05       100.00      233.85
---------------------------------------------------------------------------
Share = quantile group share of total x;
L(p)=cumulative group share; GL(p)=L(p)*mean(x)
```

- Greater dispersion (cf. quantile ratios), fall in income shares of poorer groups, but note rise in GL(p)

Picturing distributions

Kernel density estimation

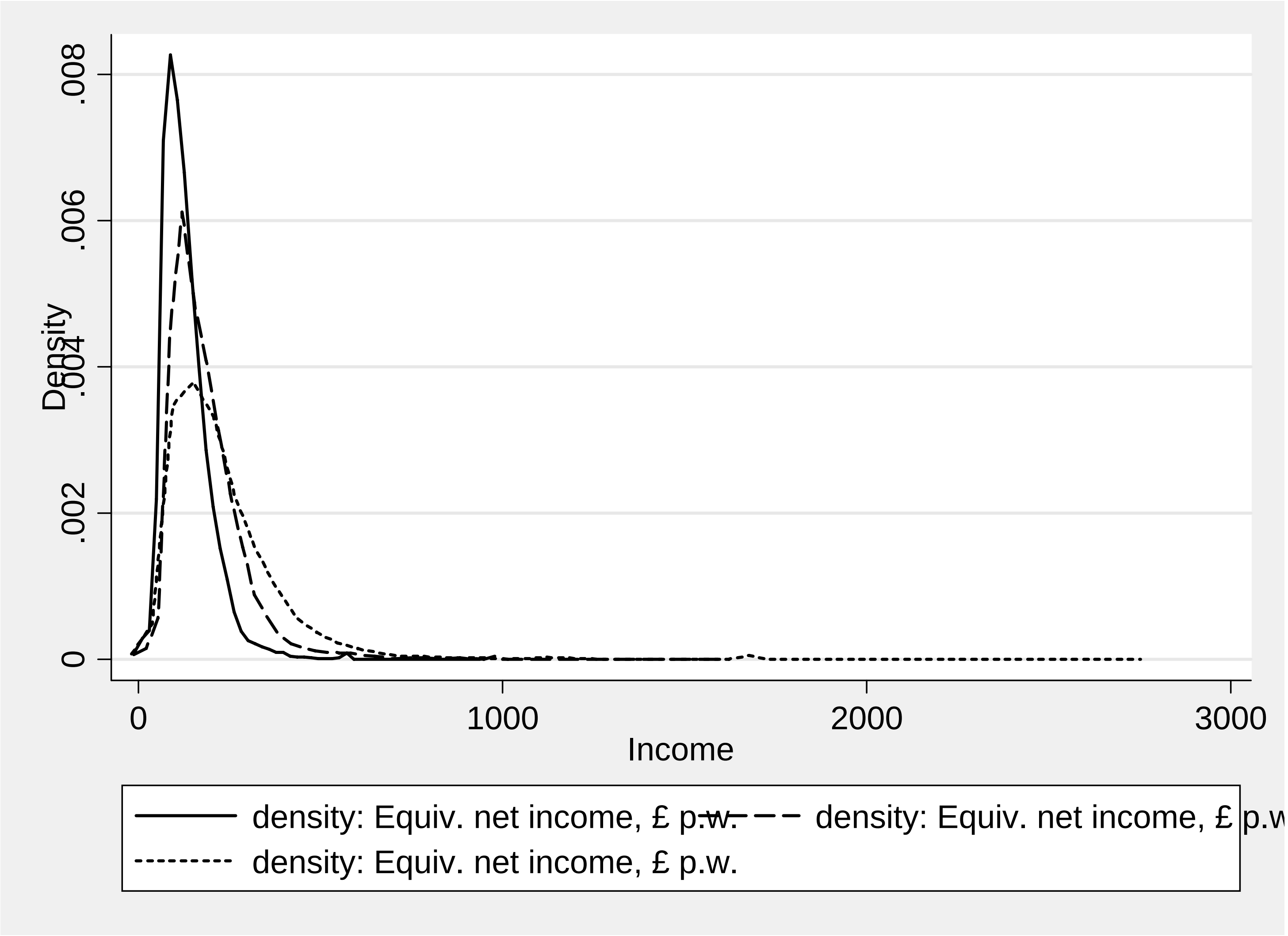
- “Smoothed histograms” are evocative of distributional shape (and have some statistical advantages compared to plain histograms)
- Highlight skewness and long right tail (of income distributions), and modality
- Can be usefully decomposed by subgroup
- But provide a ‘statistical’ description with no direct link to ‘economic’ concepts such as inequality, welfare etc

Kernel density estimation (2): `kdensity`

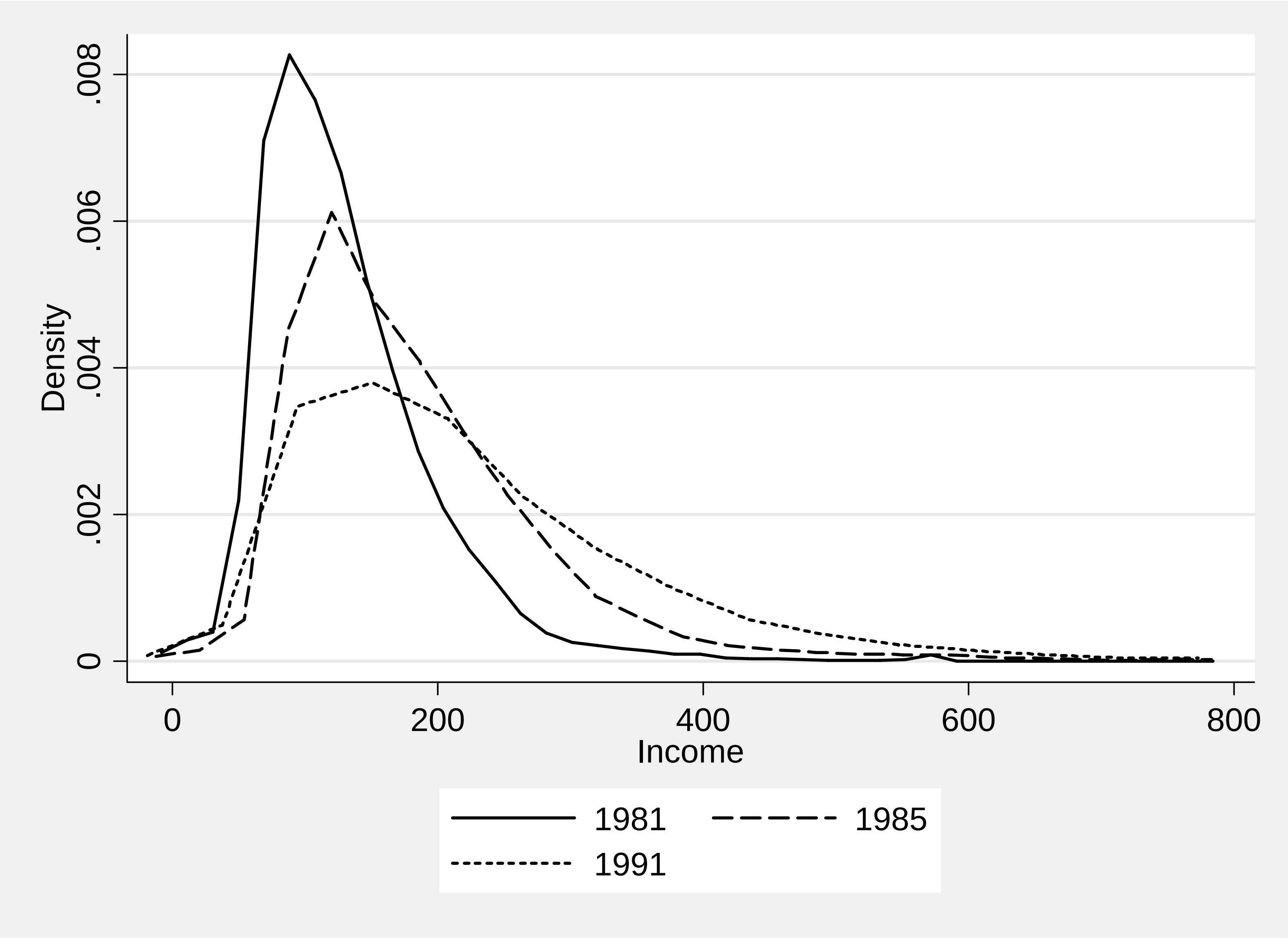
```
. * all examples use default Epanechnikov kernel, bandwidth =
   0.9*m/(nobs)^.2 where m =
   min(sqrt(var),(interquartile_range)/1.349)
. kdensity x [aw=wgt] if year == 1981, generate(x81 fx81) nograph
. kdensity x [aw=wgt] if year == 1985, generate(x85 fx85) nograph
. kdensity x [aw=wgt] if year == 1991, generate(x91 fx91) nograph
. label var x81 "1981"
. label var x85 "1985"
. label var x91 "1991"
. graph twoway (line fx81 x81, sort) (line fx85 x85, sort) (line fx91
   x91, sort) ///
   , ytitle("Density") xtitle(Income) saving(density1.gph, replace)
(file density1.gph saved)
. graph twoway (line fx81 x81 if x81 < 800, sort)(line fx85 x85 if
   x85 < 800, sort) (line fx91 x91 if x91 < 800, sort) ///
 , ytitle("Density") xtitle(Income) saving(density2.gph, replace) ///
 legend(label (1 "1981") label(2 "1985") label(3 "1991")
   region(lstyle(none)) )
(file density2.gph saved)
```

Tip: use graph command to derive basic variables, and use these as inputs to `graph twoway`

Default picture



A nicer picture



Decompositions of densities to explore the drivers of distributional change

- Overall density = population share-weighted sum of subgroup densities:

$$f(x) = \sum_{k=1}^{K} w_k \, f_k(x)$$

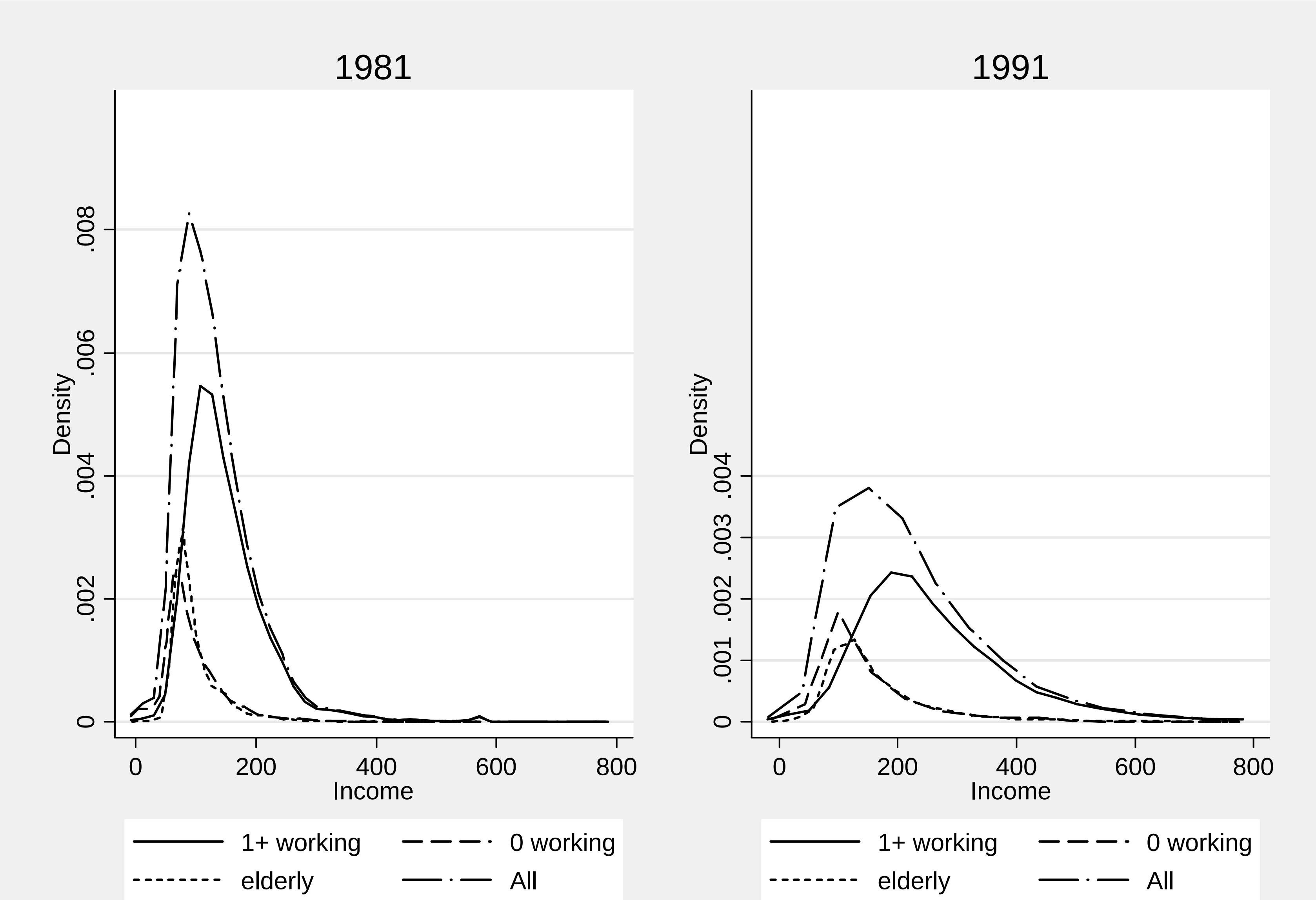
- So, change over time in density can be related to changes in subgroup population shares, w_k, and changes in the subgroup densities, $f_k(x)$
  – Allows counterfactual shift-share analysis
- Value of approach depends on judicious choice of definition of subgroup partition!
  – For sophisticated development of density decomposition methods, see Di Nardo, Fortin, Lemieux (*Econometrica* 1996); see also `dfl` on SSC. Cf. Jenkins & Van Kerm (*J Econ Inequality* 2005)

Density decomposition (2)

- Decompositions by work status useful for this period in UK given arguments about (a) the shift from work over the decade, and (b) changes in earnings distribution:
- 'Work status of family': 1 "1+ full-time worker(s)", 2 "no full-time workers", 3 "elderly (head|spouse aged 60+)"

Column %	1981	1991
1. 1+ full-time working	66.1	61.7
2. No full-time workers	17.7	20.7
3. Elderly	16.2	17.6
All	100.0	100.0

Subgroup decomposition of densities



Each graph shows $f(x)$ and $w_k f_k(x)$ for each $k = 1, 2, 3$.

Pen’s Parade of Dwarfs and a few Giants

(Jan Pen, *Income Distribution*, 1972)

- Everyone in the population is represented by a person with height proportional to income
- Line everyone up in order from shortest (poorest) to tallest (richest)
- Have the parade march past a particular spot within one hour
- What does the silhouette of the parade look like for a typical income distribution
  - a parade of dwarfs and a few giants

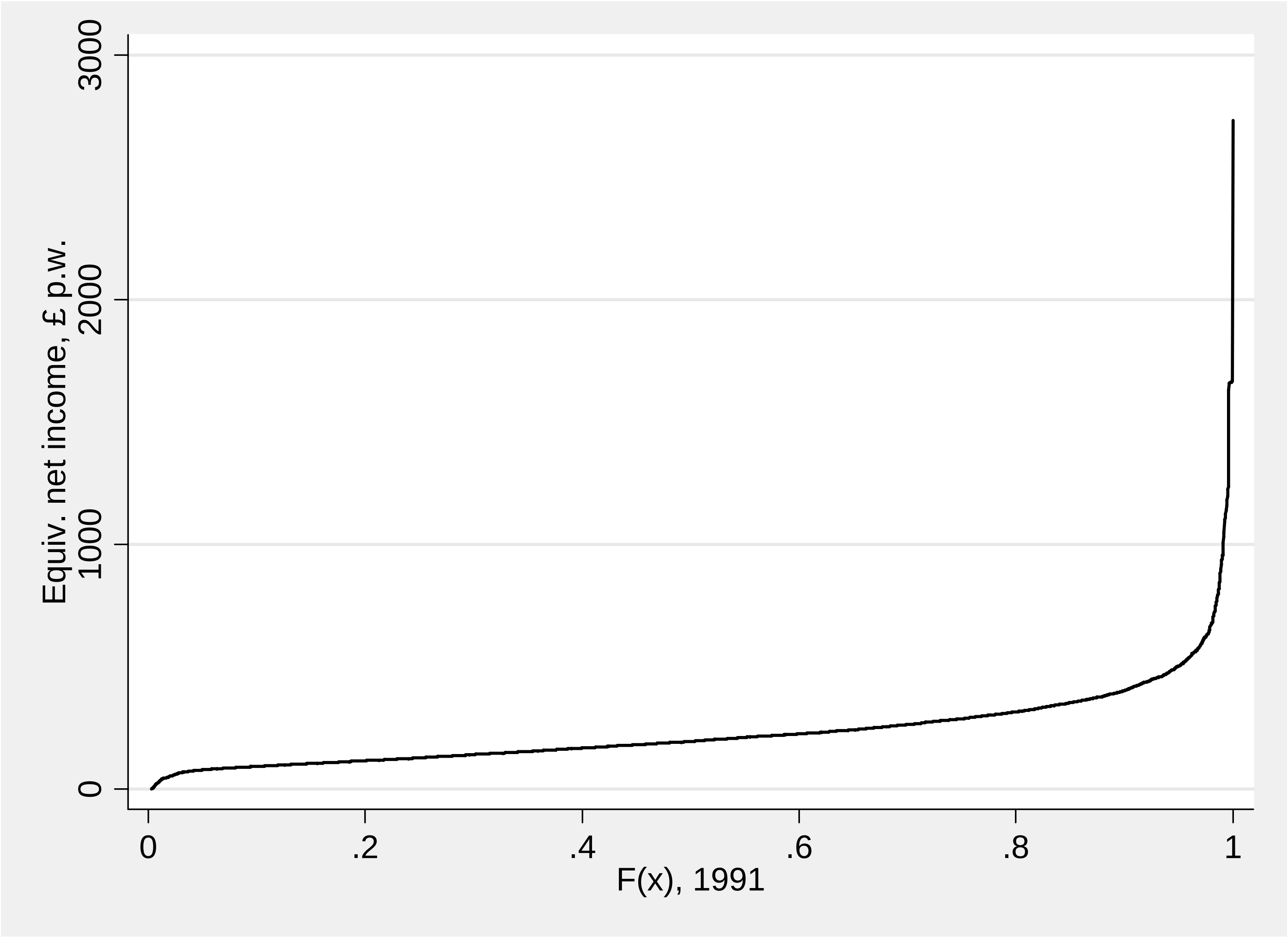
Drawing Pen's Parade

```
. cumul x  [aw = wgt] if year == 1981, generate(Fx81) equal
. cumul x  [aw = wgt] if year == 1985, generate(Fx85) equal
. cumul x  [aw = wgt] if year == 1991, generate(Fx91) equal

. lab var Fx81 "F(x), 1981"
. lab var Fx85 "F(x), 1985"
. lab var Fx91 "F(x), 1991"

. graph twoway (line x Fx91 if year==1991, sort),
   saving(parade91.gph, replace)
(file parade91.gph saved)
```

Pen’s Parade, UK 1991



- It’s just the CDF displayed differently!
- Focuses on extremes, with detail lost from middle

Pen's Parade and distributional comparisons

- Link with first order Welfare dominance
- Link with Poverty dominance

but, first, an aside about the dominance approach:

1. Characterize a class of social evaluation functions in terms of their properties
2. Show that specific configurations of particularly-defined graphs (e.g. non-intersection) are equivalent to unambiguous orderings by all social evaluation functions in the characterized class

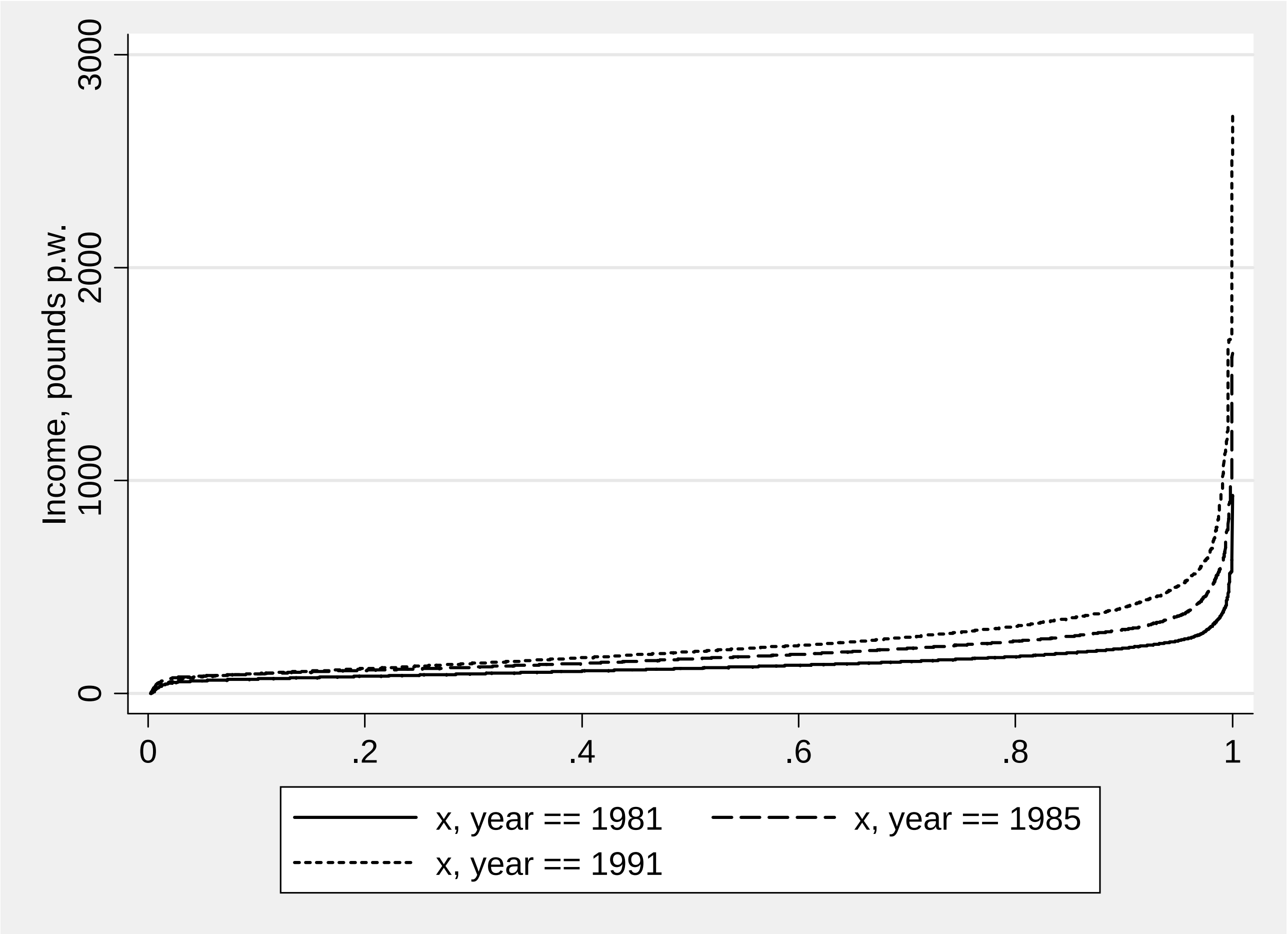
- Social welfare function $W = W(x_1,x_2,\ldots,x_n)$
  - Class $\mathcal{W}_1$ characterized by all W that are
    - increasing ($\partial W/\partial x_i > 0$, all i),
    - symmetric (invariant to permutations of the income vector)
    - replication-invariant (invariant to replications of the population)

Pen's Parade and dominance

- First order welfare dominance result (Saposnik):
  – The Parade diagram for distribution $\boldsymbol{x}$ lies everywhere above the Parade diagram for distribution $\boldsymbol{y} \Leftrightarrow W(\boldsymbol{x}) > W(\boldsymbol{y})$ for all $W \in \mathcal{W}_1$, i.e. symmetric replication-invariant social welfare functions increasing in each income
- Poverty Dominance according to the Headcount Ratio measure, H, a.k.a. proportion poor (Foster & Shorrocks)
  – If Parade diagram for distribution $\boldsymbol{x}$ lies to left of diagram for $\boldsymbol{y}$ at every income $x, y \in [0, z^*]$, then $H(\boldsymbol{x}) < H(\boldsymbol{y})$ for all common poverty lines between 0 and z^*.

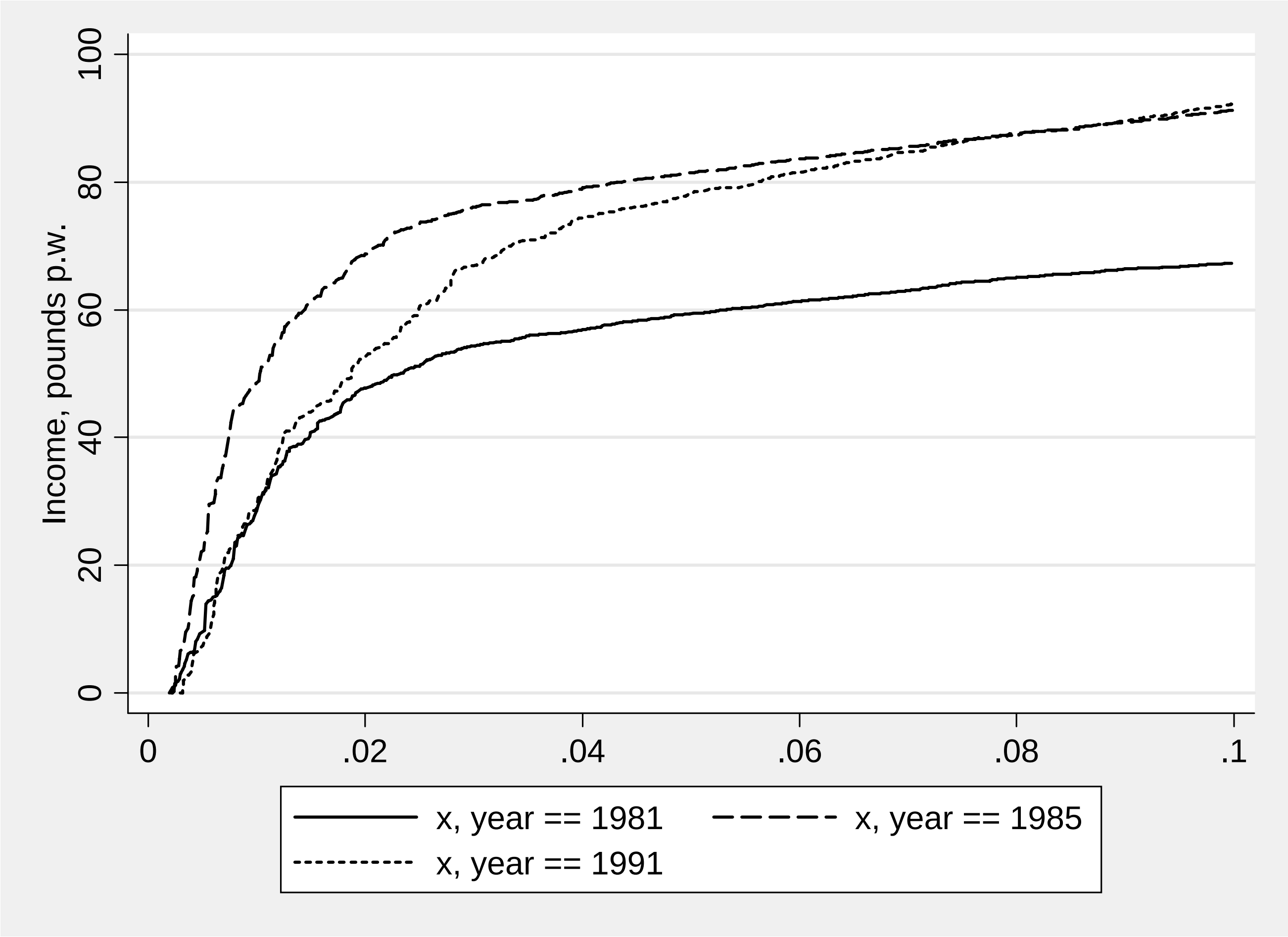
Comparing Parades over time

```
separate x, by(year)
graph twoway (line x1981 Fx81, sort) ///
        (line x1985 Fx85, sort)       ///
        (line x1991 Fx91, sort), ytitle("Income, pounds p.w.") ///
          saving(cumdist1.gph, replace)
```



Do the CDFs cross at the very bottom?

(Plot for the poorest tenth only)



General issue: the fine detail may matter. Plotting at only a limited number of selected values of *p* may mislead.

Lorenz curves and inequality

- A Lorenz curve is a plot of the cumulative income share of the poorest 100*p*% 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 $I(x_1, x_2, \ldots, x_n)$

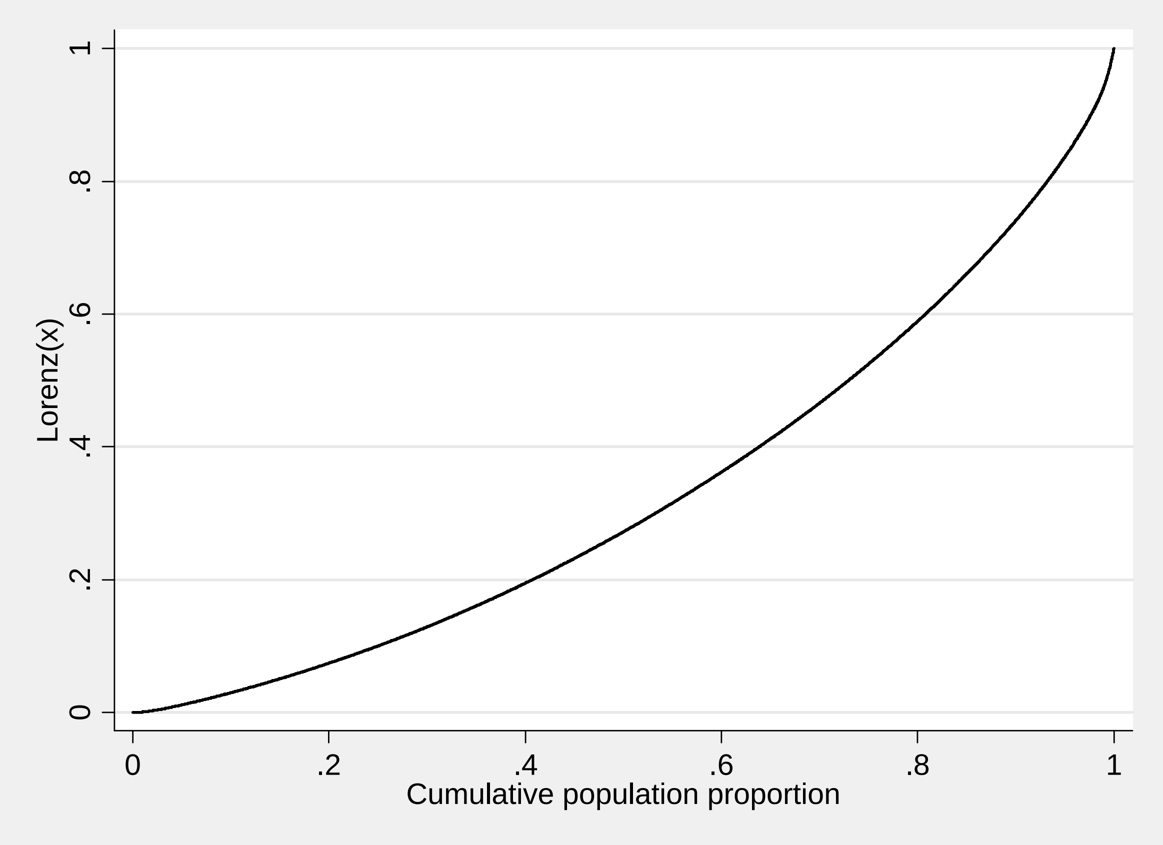
1. Symmetry a.k.a. Anonymity: only the income values matter, and no other information (permutation invariant)
2. Scale invariance: invariant to proportional scaling of all incomes
3. Replication Invariance: invariant to replications of the population
4. Principle of Transfers: a transfer of a small amount 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 $I(\boldsymbol{x}) < I(\boldsymbol{y})$

Drawing a Lorenz curve: `glcurve`

(default picture with `lorenz` option)

```
glcurve x [aw = wgt] if year == 1991,
  lorenz saving(lorenz91.gph, replace)
```

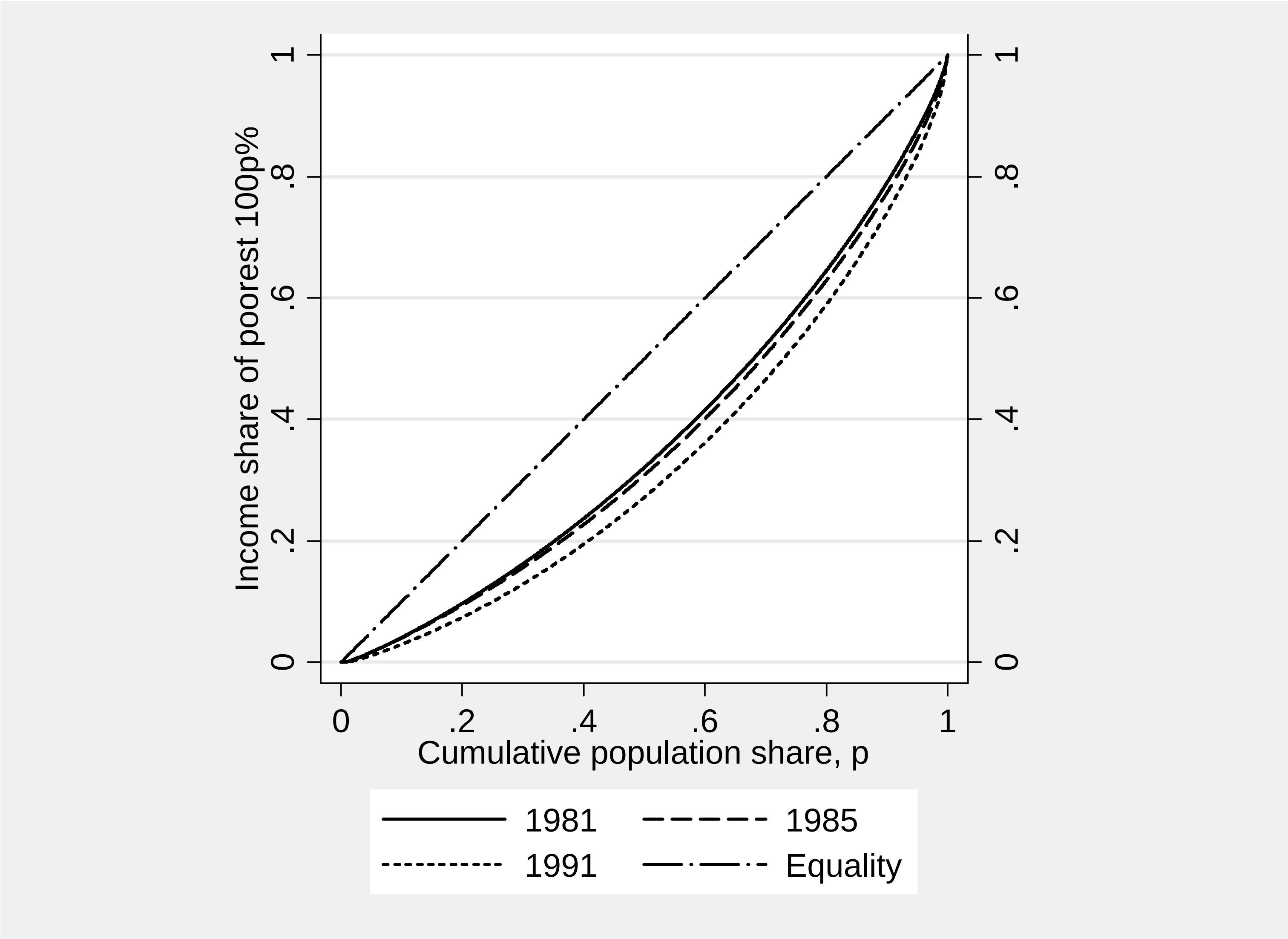


Inequality comparisons using `glcurve`

```
glcurve x [aw = wgt] , by(year) split pvar(prl) glvar(rl) lorenz
   nograph
lab var rl_1981 "1981"
lab var rl_1985 "1985"
lab var rl_1991 "1991"
lab var prl "Cumulative population share"
sort prl       // important to do this, or lines not drawn right

graph twoway (line rl_1981 prl, yaxis(1 2) )  ///
   (line rl_1985 prl, yaxis(1 2) )    ///
   (line rl_1991 prl, yaxis(1 2) )                ///
   (function y = x, range(0 1) yaxis(1 2) )       ///
   , aspect(1) xtitle("Cumulative population share, p")     ///
   ytitle("Income share of poorest 100p%", axis(1)) ytitle(" ",
   axis(2)) ///
   legend(label (1 "1981") label(2 "1985") label(3 "1991") label(4
   "Equality") ///
   region(lstyle(none)) ) saving(rl81_91, replace)
```

Inequality comparisons: 1981, 1985, 1991



Did inequality unambiguously increase over time?

What if Lorenz curves cross?

- Clear cut inequality rankings may be possible for a narrower class of inequality measures
  - Transfer Sensitivity axiom: inequality-reducing impact of a mean-preserving progressive transfer is greater the lower the income of the recipient
  - Result: if LC(x) intersects LC(y) *once* from above, then $I(\boldsymbol{x}) < I(\boldsymbol{y})$ for all inequality measures satisfying axioms 1–4 and transfer sensitivity *iff* $\mathrm{CV}(\boldsymbol{x}) < \mathrm{CV}(\boldsymbol{y})$
- You might choose to rank distributions in terms of *social welfare* rather than inequality *per se*, i.e. incorporating average living standards comparisons as well as inequality comparisons
  - cf. First Order Welfare Dominance earlier

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

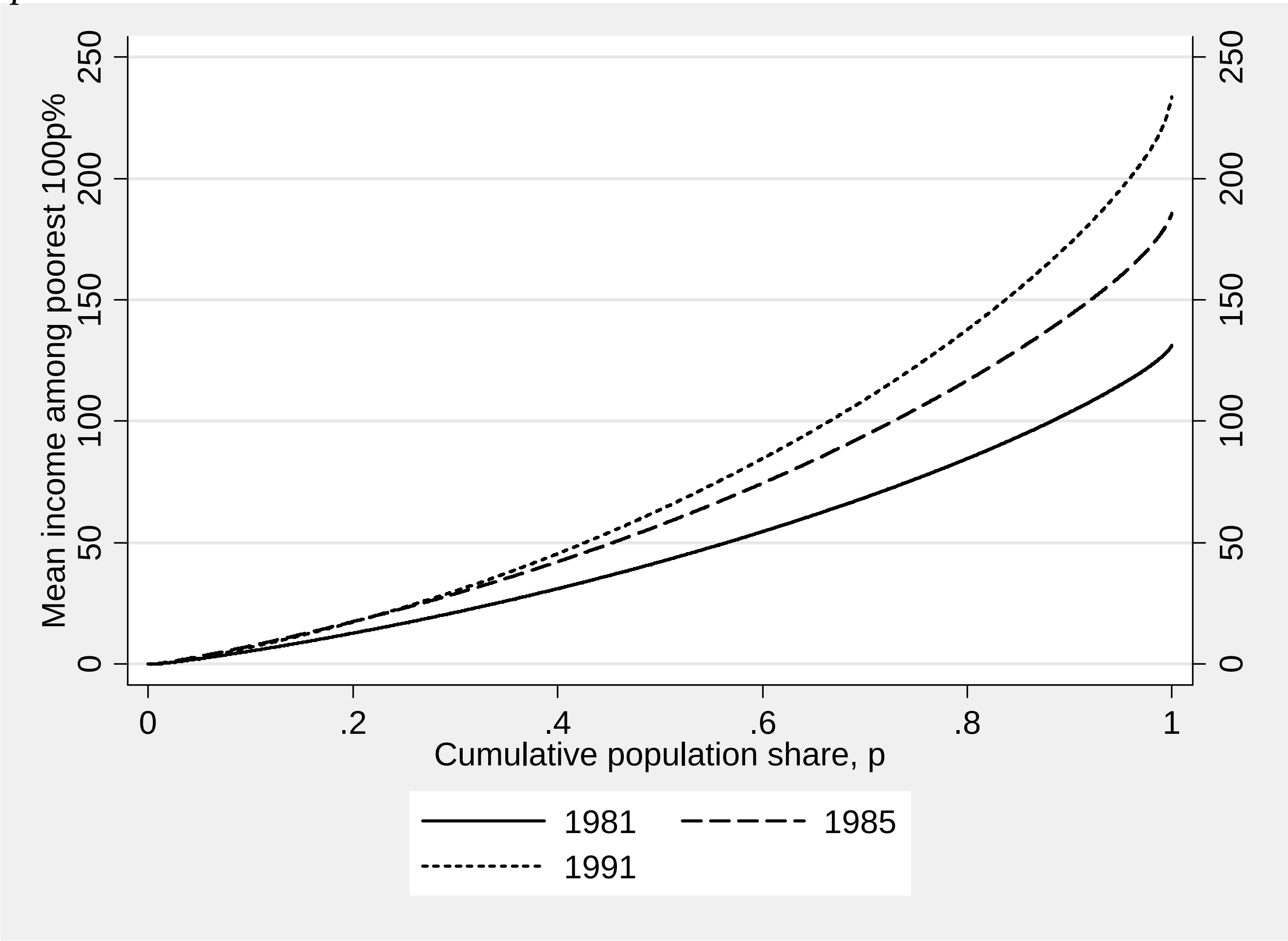
- Use `glcurve` (default graph)

```
glcurve x [aw = wgt] , by(year) split pvar(pgl) glvar(gl) nograph
lab var gl_1981 "1981"
lab var gl_1985 "1981"
lab var gl_1991 "1991"
lab var pgl "Cumulative population share, p"
sort pgl       // important to do this, or lines not drawn right

graph twoway (line gl_1981 pgl, yaxis(1 2) )        ///
       (line gl_1985 pgl, yaxis(1 2) )      ///
       (line gl_1991 pgl, yaxis(1 2) ) ,   ///
       xtitle("Cumulative population share, p") ///
       ytitle("Mean income among poorest 100p%") ///
       legend(label (1 "1981") label(2 "1985") label(3 "1991")
   region(lstyle(none)) ) ///
        saving(gl81_91, replace)
```

Generalized Lorenz curves (2)

$p\mu_p$



Overall means shown at $p = 1$

Three 'I's of Poverty (TIP) curves

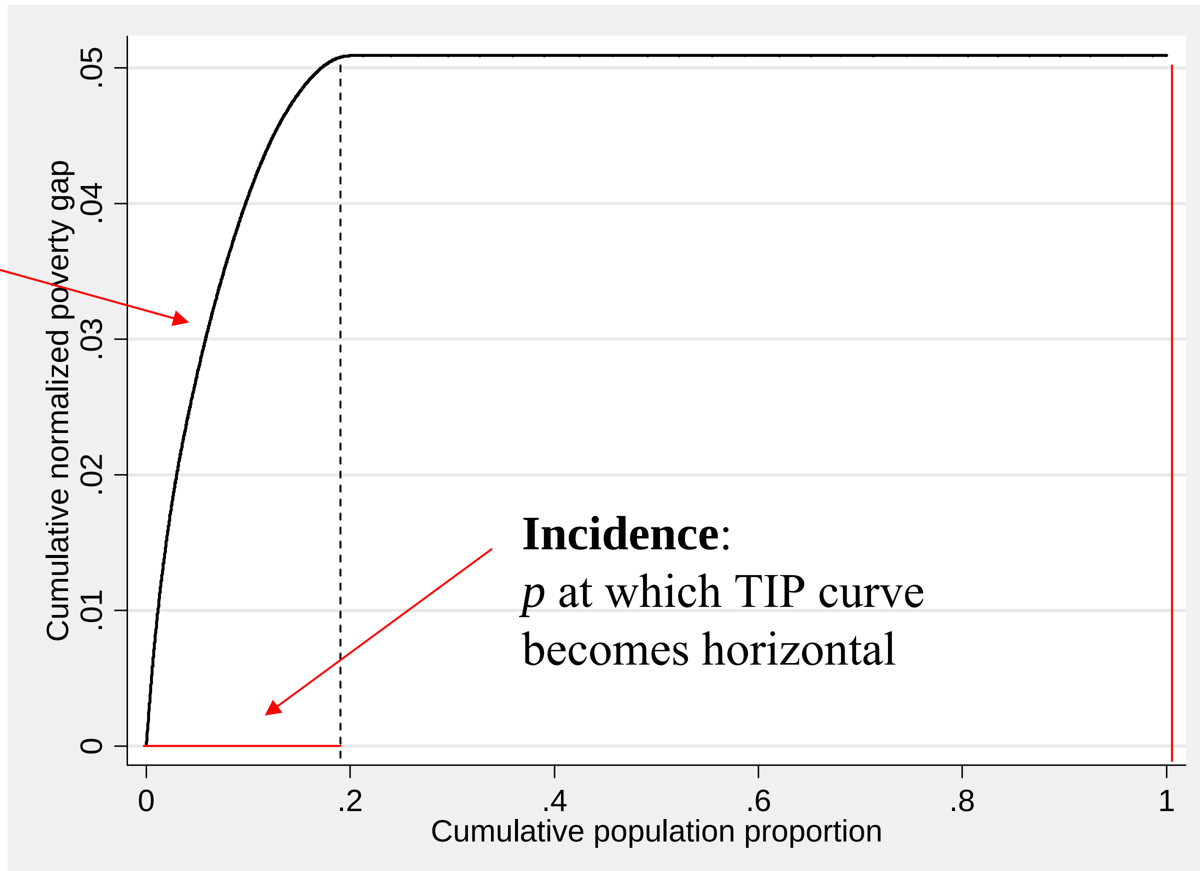
- The Three I's of Poverty (TIP):
  - Incidence: proportion poor
  - Intensity: related to average income among the poor
  - Inequality: related to the distribution of shortfalls of income from the poverty line among the poor
- The TIP curve shows the 3 'I's, and can be used for poverty dominance checks (for a given poverty line z)
- TIP curve: a plot of cumulative normalized poverty gaps against cumulative population share p, where units are ordered in ascending order of income
  - Normalized poverty gap:

$$g = (z - x)/z \qquad \text{if } x < z$$
$$g = 0 \qquad \text{otherwise}$$

Drawing a TIP curve

```
glcurve x [aw = wgt] if year == 1991, rtip(`z_91') ///
   ytitle("Cumulative normalized poverty gap") ///
   glvar(tip91) pvar(p91) saving(tip91.gph, replace)
```

Inequality: curvature



Intensity: height of TIP curve at $p = 1$

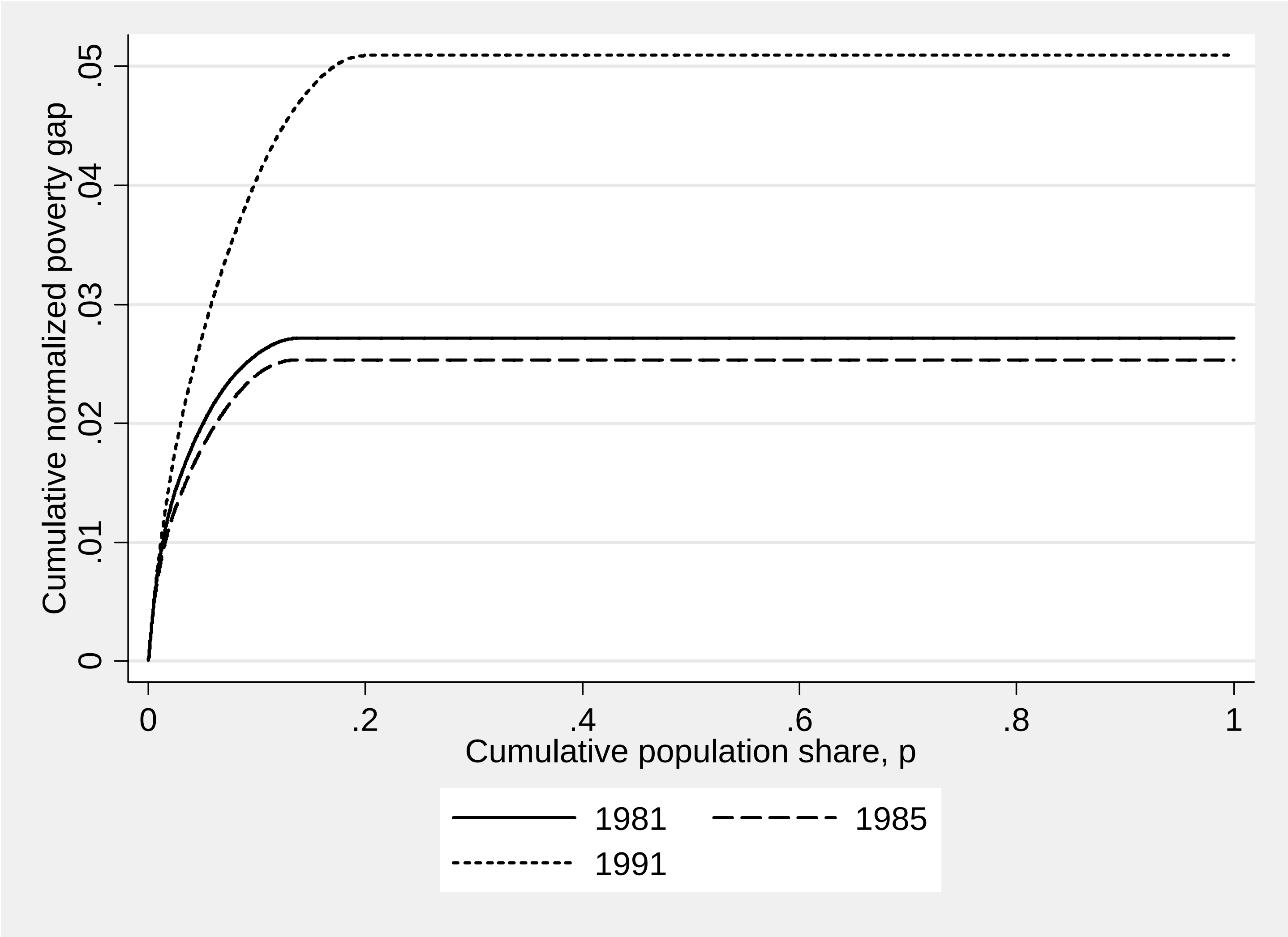
Area under TIP curve is half the SST poverty index

TIP curves and poverty dominance

- TIP dominance result (Jenkins & Lambert)

  Suppose poverty line is z. TIP($\mathbf{x}$) above TIP($\mathbf{y}$) $\Leftrightarrow$ $P(\mathbf{x}) > P(\mathbf{y})$ for all poverty indices P that are increasing, convex, replication-invariant functions of normalized poverty gap vectors g_x, g_y for all poverty lines set at z or less.
  - P includes many widely-used poverty indices, but not H (but can see H from TIP curve configuration anyway)
- Further dominance results available if TIP curves cross once, but for a restricted class of poverty indices

TIP curve comparisons using `glcurve`



Uses `graph twoway` code similar to that shown for GL curve comparisons

Usually you would only plot for $p < 0.30$ (say) in order to focus on smallest p

Summarizing distributions using parametric indices of inequality, welfare and poverty

General approach to index derivation

- Dominance approach may provide an unambiguous ordering (hence robust to differences in social judgements), but …
  - Even if there's dominance, you may want to know the *magnitude* of the difference
  - You may wish to do further analysis that is not feasible using a graphical approach, e.g. particular types of decomposition
  - There may not be dominance, and so additional judgements have to be imposed anyway in order to derive unambiguous orderings

Three related approaches to derivation and assessment of indices

1. Place additional assumptions on the social evaluation function
   – e.g. Atkinson inequality indices
2. Use a fully axiomatic approach to characterize (class of) indices
   – e.g. Generalized Entropy inequality indices
3. Continue to use ‘statistical’ indices, but also ‘reverse-engineer’ them to uncover and assess the underlying axioms and social evaluation functions
   – e.g. (generalized) Gini inequality indices
   – e.g. variance of logs (does not always satisfy the Principle of Transfers!)

Atkinson inequality indices

- Class of social welfare functions, $\mathcal{W}_2$ with $W \in \mathcal{W}_2$ if increasing in each income, symmetric, replication-invariant and concave
- Suppose also that additively separable $W = \frac{1}{N}\sum_{i=1}^{N} U(x_i)$
- Suppose $U(x_i)$ has constant elasticity. Combined with the other assumptions, this implies

$$U(x_i) = a + b\left(x_i\right)^{1-\varepsilon} / (1-\varepsilon),\ \ \varepsilon \geq 0, \varepsilon \neq 1$$

$$U(x_i) = \log(x_i),\ \ \varepsilon = 1$$

Atkinson inequality indices (2)

- Define the *equally-distributed equivalent income*, x_ε: the income which if equally distributed would produce the same level of social welfare as the original distribution

$$\frac{1}{N}\sum_{i=1}^{N} U(x_i) = \frac{1}{N}\sum_{i=1}^{N} U(x_e) = U(x_e)$$

  - NB $x_\varepsilon < \mu$, since $W \in \mathcal{W}_2$ builds in a preference for equality, other things being equal

- Inequality measure equals the 'proportionate cost of inequality',

$$A_\varepsilon = 1 - (x_\varepsilon/\mu)$$

NB $x_\varepsilon = \mu(1-A_\varepsilon)$

- So, substituting in from above …

Atkinson inequality indices (3)

$$A_\varepsilon(\mathbf{x}) = 1 - \left[\left(\frac{1}{N} \right) \sum_{i=1}^{N} (x_i / \mu)^{1-\varepsilon} \right]^{\frac{1}{1-\varepsilon}}, \ \varepsilon \geq 0, \varepsilon \neq 1$$

$$A_1(\mathbf{x}) = 1 - \exp\left[\left(\frac{1}{N} \right) \sum_{i=1}^{N} \log(x_i / \mu) \right], \ \varepsilon = 1$$

- Every member of the class satisfies inequality axioms 1–4 ('Lorenz consistent') plus Transfer Sensitivity
- ε : degree of inequality aversion
  - Larger ε means more inequality averse, or …
  - Larger ε means more sensitive to income differences at bottom of income distribution
    - With unit record household survey data, results can be very sensitive to prevalence of a few low incomes if $\varepsilon > 2$

Generalized Entropy indices

Derivation (i): strengthen the Principle of Transfers axiom (Cowell & Kuga)

- Suppose that the increase in inequality caused by a mean-preserving spread is a function of the *distance between the income shares* of the donor and recipient, and use a one-parameter distance function

Generalized Entropy indices (2)

Derivation (ii): incorporate an additional axiom, notably additive decomposability by population subgroup (Cowell, Bourguignon, Shorrocks):

- Total inequality = weighted sum of the inequalities within each subgroup, plus inequality between groups

$$I(\boldsymbol{x}) = I_{Within} + I_{Between}$$

where $I_{Within} = \Sigma_k w_k I(\boldsymbol{x}_k)$ for subgroups $k = 1,\ldots, K$

$I_{Between} = I(\mu_1, \mu_2, \ldots, \mu_k)$

$w_k = w_k(\mu_k, N_k)$

Useful for counterfactual shift-share analysis of inequality trends (subject to good choice of groups!)

Generalized Entropy indices (3)

The combination of the axioms implies:

$$I_a(\mathbf{x}) = \left(\frac{1}{a(a-1)}\right)\left[\left(\left(\frac{1}{N}\right)\sum_{i=1}^{N}(x_i/\mu)^a\right)-1\right],\ a \neq 0,1$$

$$I_1(\mathbf{x}) = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(x_i/\mu)\log(x_i/\mu),\ \ a=1$$

$$I_0(\mathbf{x}) = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\log(\mu/x_i),\ \ a=0$$

I_2 is half CV squared; I_1 is Theil index; I_0 is Mean Log Deviation

Subgroup aggregating weight w_k is subgroup population share for I_0, subgroup income share for I_1. For other GE indices, weights do not sum to one across K.

Generalized Entropy indices (4)

- Parameter a specifies sensitivity to income differences in different parts of the income distribution
- Larger $a > 0$ corresponds to greater sensitivity to high income values; smaller $a < 0$, greater sensitivity to low income values
  - With unit record household survey data, results can be sensitive to the prevalence of a few high incomes if $a \geq 2$ (beware the top-sensitive CV!) and the prevalence of a few tiny incomes if $a \leq -1$
  - MLD is relatively 'middle sensitive' ; cf. Gini coefficient (most sensitive to transfers round the mode)
  - I_a is Transfer Sensitive if $a < 2$
- For every member of the Atkinson class A_ε, there is an ordinally equivalent member of the GE class $I_{1-\varepsilon}$
- GE class is additively decomposable, but the Atkinson class is not (it's decomposable in another sense), and the Gini coefficient is not decomposable in either sense

Calculating inequality indices using `ineqdeco`, `ineqdec0`

- Indices calculated:
  - $p90/p10$, $p75/p25$
  - Gini
  - Generalized Entropy, $a = -1, 0, 1, 2$
  - Atkinson , $\varepsilon = 0.5, 1, 2$, plus
  - optional decompositions by population subgroup, and
  - optional selected social welfare indices (see help files)
- By contrast with `ineqdeco`, `ineqdec0` allows zero and negative values, but only reports results for subset of indices (percentile ratios, I_2, Gini)
  - may be more useful for analyzing wealth distributions

ineqdeco (1981)

```
. ineqdeco x [aw = wgt] if year == 1981

Warning: x has 20 values = 0. Not used in calculations

Percentile ratios for distribution of x: all valid obs.
------------------------------------------------------------
p90/p10  p90/p50  p10/p50  p75/p25  p75/p50  p25/p50
------------------------------------------------------------
  3.131    1.804    0.576    1.869    1.369    0.733
Generalized Entropy indices GE(a), where a = income difference
 sensitivity parameter, and Gini coefficient
----------------------------------------------------------------------
  All obs |     GE(-1)       GE(0)       GE(1)       GE(2)        Gini
----------+-----------------------------------------------------------
          |    0.19021     0.11429     0.11169     0.12879     0.25739
----------------------------------------------------------------------
Atkinson indices, A(e), where e > 0 is the inequality aversion parameter
----------------------------------------------
  All obs |     A(0.5)        A(1)        A(2)
----------+-----------------------------------
          |    0.05432     0.10800     0.27558
----------------------------------------------
```

ineqdeco (1991)

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

Warning: x has 20 values = 0. Not used in calculations

Percentile ratios for distribution of x: all valid obs.
------------------------------------------------------------
p90/p10  p90/p50  p10/p50  p75/p25  p75/p50  p25/p50
------------------------------------------------------------
  4.336    2.063    0.476    2.249    1.474    0.655
Generalized Entropy indices GE(a), where a = income difference
 sensitivity parameter, and Gini coefficient
----------------------------------------------------------------------
  All obs |     GE(-1)       GE(0)       GE(1)       GE(2)       Gini
----------+-----------------------------------------------------------
          |    3.68289     0.19524     0.20039     0.27432    0.33465
----------------------------------------------------------------------
Atkinson indices, A(e), where e > 0 is the inequality aversion parameter
----------------------------------------------
  All obs |     A(0.5)        A(1)        A(2)
----------+-----------------------------------
          |    0.09294     0.17736     0.88047
----------------------------------------------
```

All indices show a rise, but note extraordinary rise in the most bottom-sensitive indices and (to lesser extent) top-sensitive ones. See earlier data checks!!

Decomposition by population subgroup

(by work status, as defined earlier)

```
. ineqdeco x [aw = wgt] if year == 1991, by(wkstatus)
< … output omitted … i.e. estimates of overall inequality >
Subgroup summary statistics, for each subgroup k = 1,...,K:
-----------------------------------------------------------------------------------------------
        family work status |   Pop. share          Mean      Rel.mean  Income share      log(mean)
---------------------------+-------------------------------------------------------------------
1+ full-time, non-elderly |      0.61724     278.80399       1.18873       0.73373       5.63051
 0 full-time, non-elderly |      0.20607     151.63173       0.64651       0.13323       5.02145
     head|spouse aged 60+ |      0.17669     176.60454       0.75298       0.13304       5.17391
-----------------------------------------------------------------------------------------------
Subgroup indices: GE_k(a) and Gini_k
-----------------------------------------------------------------------------------------
        family work status |     GE(-1)       GE(0)       GE(1)       GE(2)        Gini
---------------------------+-------------------------------------------------------------
1+ full-time, non-elderly |    0.23007     0.15369     0.15986     0.21307     0.29524
 0 full-time, non-elderly |   10.92318     0.18368     0.18899     0.28463     0.31911
     head|spouse aged 60+ |    0.19322     0.16537     0.20239     0.34651     0.31284
-----------------------------------------------------------------------------------------
Within-group inequality, GE_W(a)
-----------------------------------------------------------
  All obs |     GE(-1)       GE(0)       GE(1)       GE(2)
----------+------------------------------------------------
          |    3.64657     0.16194     0.16940     0.24507
-----------------------------------------------------------
Between-group inequality, GE_B(a):
-----------------------------------------------------------
  All obs |     GE(-1)       GE(0)       GE(1)       GE(2)
----------+------------------------------------------------
          |    0.03632     0.03330     0.03099     0.02926
```

In which group is the tiny income with a very large influence on calculations?

Decomposition output (continued)

Subgroup Atkinson indices, A_k(e)

family work status	A(0.5)	A(1)	A(2)
1+ full-time, non-elderly	0.07438	0.14246	0.31513
0 full-time, non-elderly	0.08628	0.16780	0.95623
head|spouse aged 60+	0.08630	0.15242	0.27872

Within-group inequality, A_W(e)

All obs	A(0.5)	A(1)	A(2)
	0.07755	0.14716	0.39570

Between-group inequality, A_B(e)

All obs	A(0.5)	A(1)	A(2)
	0.01668	0.03541	0.80219

Selected welfare indices ("w" option)

- Equally-distributed-equivalent incomes, $\varepsilon = 0.5, 1, 2$
- Social welfare indices, $\varepsilon = 0.5, 1, 2$
  - Both types are 'Generalized Lorenz' consistent
- Sen's welfare index = mean*(1 – Gini)

```
Equally-distributed-equivalent incomes, Yede(e)
----------------------------------------------
  All obs |  Yede(0.5)     Yede(1)     Yede(2)
----------+-----------------------------------
          |  212.74117   192.94182    28.03559
----------------------------------------------

Social welfare indices, W(e), and Sen's welfare index
----------------------------------------------------------------------
  All obs |        W(0.5)           W(1)           W(2)  mean*(1-Gini)
----------+-----------------------------------------------------------
          |      29.17130        5.26239       -0.03567      156.05151
----------------------------------------------------------------------
```

Calculating poverty indices with povdeco

- Indices calculated:
  - FGT_0 = headcount ratio = proportion poor
  - FGT_1 = averaged normalized poverty gap
  - FGT_2 = averaged normalized squared poverty gap

$$FGT_\alpha(\mathbf{x}; z) = \left(\frac{1}{N}\right)\sum_{i=1}^{N} I(x < z)\left[1 - (x/z)\right]^\alpha, \ \alpha > 0$$

where α is a 'poverty aversion' parameter (larger α gives greater weight to larger poverty gaps, i.e. poorer people)

  - plus various auxiliary information
  - plus optional decomposition by population subgroup:

$$FGT_\alpha(\mathbf{x}; z) = \sum_{k=1}^{K} \left(N_k / N\right) FGT_\alpha(\mathbf{x}_k; z), \ \alpha > 0$$

povdeco (1981 and 1991)

```
. povdeco x [aw = wgt] if year == 1981,
    pline(`z_81')

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

Total number of observations = 9772
Weighted total no. of observations = 54872650
Number of observations poor = 1248
Weighted no. of obs poor = 7737729
Mean of x amongst the poor =    58.411
Mean of poverty gaps (poverty line - x) amongst
    the poor =    13.933

Foster-Greer-Thorbecke poverty indices, FGT(a)

----------------------------------------------
  All obs |       a=0         a=1         a=2
----------+-----------------------------------
          |    0.14101     0.02716     0.01174
----------------------------------------------
FGT(0): headcount ratio (proportion poor)
FGT(1): average normalised poverty gap
FGT(2): average squared normalised poverty gap
```

```
. povdeco x [aw = wgt] if year == 1991,
    pline(`z_91')

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

Total number of observations = 6468
Weighted total no. of observations = 55851705
Number of observations poor = 1322
Weighted no. of obs poor = 11289372
Mean of x amongst the poor =    86.947
Mean of poverty gaps (poverty line - x) amongst
    the poor =    29.279

Foster-Greer-Thorbecke poverty indices, FGT(a)

----------------------------------------------
  All obs |       a=0         a=1         a=2
----------+-----------------------------------
          |    0.20213     0.05092     0.02215
----------------------------------------------
FGT(0): headcount ratio (proportion poor)
FGT(1): average normalised poverty gap
FGT(2): average squared normalised poverty gap
```

Summary statistics for subgroup k = 1,...,K

family work status	Pop. share	Mean	Mean|poor	Mean gap|poor
1+ full-time, non-elderly	0.61719	278.00793	80.85588	35.37053
0 full-time, non-elderly	0.20664	150.77277	85.01402	31.21239
head|spouse aged 60+	0.17617	176.60454	94.44805	21.77834

Subgroup FGT index estimates, FGT(a)

family work status	a=0	a=1	a=2
1+ full-time, non-elderly	0.06793	0.02067	0.01201
0 full-time, non-elderly	0.48543	0.13036	0.05450
head|spouse aged 60+	0.34000	0.06371	0.01972

Subgroup poverty 'share', S_k = v_k.FGT_k(a)/FGT(a)

family work status	a=0	a=1	a=2
1+ full-time, non-elderly	0.20741	0.25056	0.33468
0 full-time, non-elderly	0.49626	0.52902	0.50846
head|spouse aged 60+	0.29633	0.22042	0.15686

Subgroup poverty 'risk' = FGT_k(a)/FGT(a) = S_k/v_k

family work status	a=0	a=1	a=2
1+ full-time, non-elderly	0.33605	0.40597	0.54226
0 full-time, non-elderly	2.40155	2.56011	2.46061
head|spouse aged 60+	1.68210	1.25117	0.89038

Decomposition of poverty by work status, 1991

(aggregate output not shown)

Variance estimation

Background

- Estimation using sample survey data means that estimates reflect sampling variability (SEs!)
- 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?
    - See Checklist earlier
  - 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 (bootstrap methods at end)

Overview

- Most poverty indices, given fixed (non-stochastic) poverty line can be expressed as means of particular variables
  - Can use Stata's `svy` commands directly or adapt them
- *However*, how to extend derivations to statistics that are not simple functions of totals?
  - GE and Atkinson inequality measures (non-linear functions of multiple moments)
  - Functions of order statistics (e.g. Gini, Lorenz curve)
  - Poverty indices, with poverty lines derived from the distribution (e.g. 60% of median) [not considered here!]
- *Answer*: linearization methods can be adapted

Assumptions about survey design

- All of the built-in and user-written programs used below have options to account for the impact of clustering and stratification

- 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
- So, we can compare variances estimated assuming SRS versus accounting for within-household clustering

Headcount ratio (with given poverty line)

- Poverty status is 0/1 variable; H = mean of this
- First you must `svyset` the data

```
 * SRS, but accounting for the weights
. svyset [pweight = wgt]

      pweight: wgt
          VCE: linearized
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>

* account for clustering within HHs
. svyset hrn [pweight = wgt]

      pweight: wgt
          VCE: linearized
     Strata 1: <one>
         SU 1: hrn
        FPC 1: <zero>
```

1981 versus 1991, assuming SRS

Independent samples, so OK to use -if- here!

```
. svy: mean poor if year == 1981
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =       1          Number of obs     =     9772
Number of PSUs   =    9772          Population size   = 5.5e+07
                                    Design df         =     9771
--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        poor |   .1410125   .0041339      .1329092    .1491158
--------------------------------------------------------------
. svy: mean poor if year == 1991
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =       1          Number of obs     =     6468
Number of PSUs   =    6468          Population size   = 5.6e+07
                                    Design df         =     6467
--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        poor |   .2021312   .0057211       .190916    .2133464
```

1981 versus 1991, accounting for HH clustering

```
. svy: mean poor if year == 1981
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =        1          Number of obs     =     9772
Number of PSUs   =     7476          Population size   = 5.5e+07
                                     Design df         =     7475
--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        poor |   .1410125   .0044859       .132219     .149806
--------------------------------------------------------------
. svy: mean poor if year == 1991
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =        1          Number of obs     =     6468
Number of PSUs   =     5254          Population size   = 5.6e+07
                                     Design df         =     5253
--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        poor |   .2021312   .0062077      .1899615    .2143009
--------------------------------------------------------------
```

Accounting for HH level clustering raises SEs, but not by a large amount

FGT(1), given poverty line, HH clustering

```
. ge ngap = poor*($z_81- x)/$z_81 if year == 1981
(15459 missing values generated)

. replace ngap = poor*($z_91 - x)/$z_91 if year == 1991
(6468 real changes made)

. svy: mean ngap if year == 1991
(running mean on estimation sample)

Survey: Mean estimation
Number of strata =        1           Number of obs     =    6468
Number of PSUs   =     5254           Population size   = 5.6e+07
                                      Design df         =    5253
--------------------------------------------------------------
             |             Linearized
             |       Mean    Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        ngap |     .05092    .0021571      .0466912    .0551488
--------------------------------------------------------------
```

First generate the unit-level poverty variable, and then take the (svy) mean of that

Linearization again: inequality indices

(Biewen & Jenkins, *OBES*, 2006)

- Each member of the GE and Atkinson classes of inequality indices can be written as function of several totals, but those totals involve several moments of the distribution

Replacing totals T by their estimates $\hat{T}$, inequality index I is then estimated as $\hat{I} = f(\hat{T})$ with

$$\hat{T}_k = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} w_{hij} t_{hijk}. \qquad (11)$$

Vector of totals

Summation over strata, clusters, units in clusters

Now apply the linearization idea …

Linearization again: inequality indices (2)

Assuming that the sample is sufficiently large that a first-order Taylor approximation of $f(\cdot)$ holds,[4] i.e.

$$f(\hat{T}) \approx f(T) + \left[\sum_{k=1}^{K} \frac{\partial f(T)}{\partial T_k}(\hat{T}_k - T_k)\right], \tag{12}$$

the variance of $\hat{I}$ can be approximated by the variance of its first-order residual

$$\sum_{k=1}^{K} \left(\frac{\partial f(T)}{\partial T_k}\right) \hat{T}_k. \tag{13}$$

As observed by Woodruff (1971), this variance can be easily determined by reversing the order of summation in the residual, i.e.

$$\mathrm{var}(\hat{I}) \approx \mathrm{var}\left(\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_i} w_{hij}\left[\sum_{k=1}^{K}\left(\frac{\partial f(T)}{\partial T_k}\right) t_{hijk}\right]\right) = \mathrm{var}(\hat{S}). \tag{14}$$

Application of linearization methods requires derivation of a “pseudo-variable” a.k.a. “first-order residual”. Complicated for inequality indices; Woodruff result helps!

Linearization again: inequality indices (3)

$$\widehat{\mathrm{var}}(\hat{I}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\sum_{j=1}^{m_i} w_{hij}\tilde{s}_{hij} - \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{m_i} w_{hij}\tilde{s}_{hij}}{n_h} \right)^2, \qquad (15)$$

with

$$\tilde{s}_{hij} = \sum_{k=1}^{K} \left(\frac{\partial f(\hat{T})}{\partial \hat{T}_k} \right) t_{hijk}.$$

Require sampling variance of a total estimator; once that found, then can use svy

$$\tilde{s}_{hij}^{\mathrm{GE}} = \frac{1}{\alpha} \hat{U}_\alpha \hat{U}_1^{-\alpha} \hat{U}_0^{\alpha-2} - \frac{1}{\alpha - 1} \hat{U}_\alpha \hat{U}_1^{-\alpha-1} \hat{U}_0^{\alpha-1} \cdot y_{hij} + \frac{1}{\alpha^2 - \alpha} \hat{U}_0^{\alpha-1} \hat{U}_1^{-\alpha} \cdot (y_{hij})^\alpha \qquad (16)$$

$$\tilde{s}_{hij}^{\mathrm{Theil}} = \hat{U}_1^{-1} \cdot y_{hij} \log y_{hij} - \hat{U}_1^{-1} (\hat{T}_{1,1} \hat{U}_1^{-1} + 1) \cdot y_{hij} + \hat{U}_0^{-1} \qquad (17)$$

$$\tilde{s}_{hij}^{\mathrm{MLD}} = -\hat{U}_0^{-1} \cdot \log y_{hij} + \hat{U}_1^{-1} \cdot y_{hij} + U_0^{-1} (\hat{T}_{0,1} \hat{U}_0^{-1} - 1) \qquad (18)$$

Here are the pseudo-variables for GE indices; analogous approach used for Atkinson indices

Linearization again: inequality indices (4)

- Estimate sampling variance for each index by calculating the relevant pseudo-variable, and calculating its approximate variance using standard methods for the variance of a total
  - `svyatk` and `svygei` (version 8 programs; `svyset` differs in v. 9)
- Related methods can be used to derive the sampling variance of the Gini index, and Lorenz ordinates and income shares
  - `svylorenz` implements formulae from Kovačević & Binder (*JOS*, 1997)

Variance estimation for GE indices, 1991

```
. * account for clustering within HHs
. version 8: svyset [pweight = wgt], psu(hrn)
pweight is wgt
psu is hrn

. svygei x if year == 1991

Warning: x has 20 values = 0. Not used in calculations

Complex survey estimates of Generalized Entropy inequality indices
pweight: wgt                                  Number of obs    = 6448
Strata: <one>                                 Number of strata = 1
PSU: hrn                                      Number of PSUs   = 5237
                                              Population size  = 55687900
---------------------------------------------------------------------------
Index     |  Estimate   Std. Err.      z      P>|z|     [95% Conf. Interval]
----------+----------------------------------------------------------------
GE(-1)    |  3.682893   3.4001584     1.08    0.279     -2.981295   10.34708
MLD       |  .1952363   .00646194    30.21    0.000      .1825711   .2079015
Theil     |  .2003897   .00793043    25.27    0.000      .1848464   .2159331
GE(2)     |   .274325   .01669517    16.43    0.000       .241603   .3070469
GE(3)     |  .5247535   .05055911    10.38    0.000      .4256594   .6238475
---------------------------------------------------------------------------
```

Relatively small "z" related to low income outlier (see earlier)

Variance estimation for *A* indices, 1991

```
. svyatk x if year == 1991

Warning: x has 20 values = 0. Not used in calculations

Complex survey estimates of Atkinson inequality indices

pweight: wgt                                  Number of obs     = 6448
Strata: <one>                                 Number of strata  = 1
PSU: hrn                                      Number of PSUs    = 5237
                                              Population size   = 55687900
---------------------------------------------------------------------------
Index     |  Estimate    Std. Err.      z       P>|z|     [95% Conf. Interval]
----------+----------------------------------------------------------------
A(0.5)    |  .0929418   .00307338    30.24    0.000      .0869181   .0989656
A(1)      |  .1773597   .00531586    33.36    0.000      .1669408   .1877786
A(1.5)    |  .3052262   .04517203     6.76    0.000      .2166907   .3937618
A(2)      |  .8804655   .0971663     9.06    0.000        .690023   1.070908
A(2.5)    |  .9911087   .00591167   167.65    0.000       .979522   1.002695
---------------------------------------------------------------------------
 Excluding income x < 1
 A(2.5)   |  .5164274   .03901169    13.24    0.000      .4399659   .5928889
```

Estimates for a subgroup (subpop option)

```
. ta but if year == 1991, ge(bu)
benefit unit type |      Freq.     Percent        Cum.
------------------+-----------------------------------
 couple pensioner |        579        8.95        8.95
 single pensioner |      1,050       16.23       25.19
couple with child |      1,371       21.20       46.38
  couple no child |      1,303       20.15       66.53
single with child |        282        4.36       70.89
  single no child |      1,883       29.11      100.00
------------------+-----------------------------------
            Total |      6,468      100.00

. svygei x if year == 1991, subpop(bu5)

Warning: x has 20 values = 0. Not used in calculations

Complex survey estimates of Generalized Entropy inequality indices

pweight: wgt                                  Number of obs    = 6448
Strata: <one>                                 Number of strata = 1
PSU: hrn                                      Number of PSUs   = 5237
                                              Population size  = 55687900
Subpop: bu5, subpop. size = 3517058
---------------------------------------------------------------------------
Index     |  Estimate   Std. Err.       z       P>|z|     [95% Conf. Interval]
---------+-----------------------------------------------------------------
GE(-1)    |  .1148695  .01864548     6.16    0.000      .078325   .1514139
MLD       |  .1023959  .0110642     9.25    0.000     .0807104   .1240813
Theil     |   .109235  .01211802     9.01    0.000     .0854841   .1329858
GE(2)     |  .1318947  .0176408     7.48    0.000     .0973194   .1664701
GE(3)     |  .1801924  .03148777     5.72    0.000     .1184775   .2419073
---------------------------------------------------------------------------
```

Variance estimation for shares, Lorenz curve and Gini: svylorenz

```
. svylorenz x if year == 1991
Warning: x has 20 values = 0. Used in calculations

Variance estimation: quantile group shares and cumulative shares, and Gini
Number of strata =          1               Number of obs     =          6468
Number of PSUs   =       5254               Population size   =   55851705.00
                                            Design df         =          5253
---------------------------------------------------------------------------
  Group  |            Linearized
  share  |  Estimate   Std. Err.        z      P>|z|     [95% Conf. Interval]
---------+-----------------------------------------------------------------
     1   |  0.029606   0.010052     2.945    0.003      .0099037   .0493085
     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    40.285    0.000       .246449    .271657
```

Variance estimation for shares, Lorenz curve and Gini (continued)

Cumul. share						
1	0.029606	0.010052	2.945	0.003	.0099037	.0493085
2	0.074109	0.009867	7.511	0.000	.0547691	.0934483
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	58.876	0.000	.450593	.481626
8	0.589496	0.007274	81.037	0.000	.575238	.603753
9	0.740947	0.006431	115.223	0.000	.728343	.753551
10	1.000000					
Gini	.3365993	.00515134	65.342	0.000	.3265028	.3466957

Gini calculations are based on the complete unit record data

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

Lorenz curve comparisons with CIs

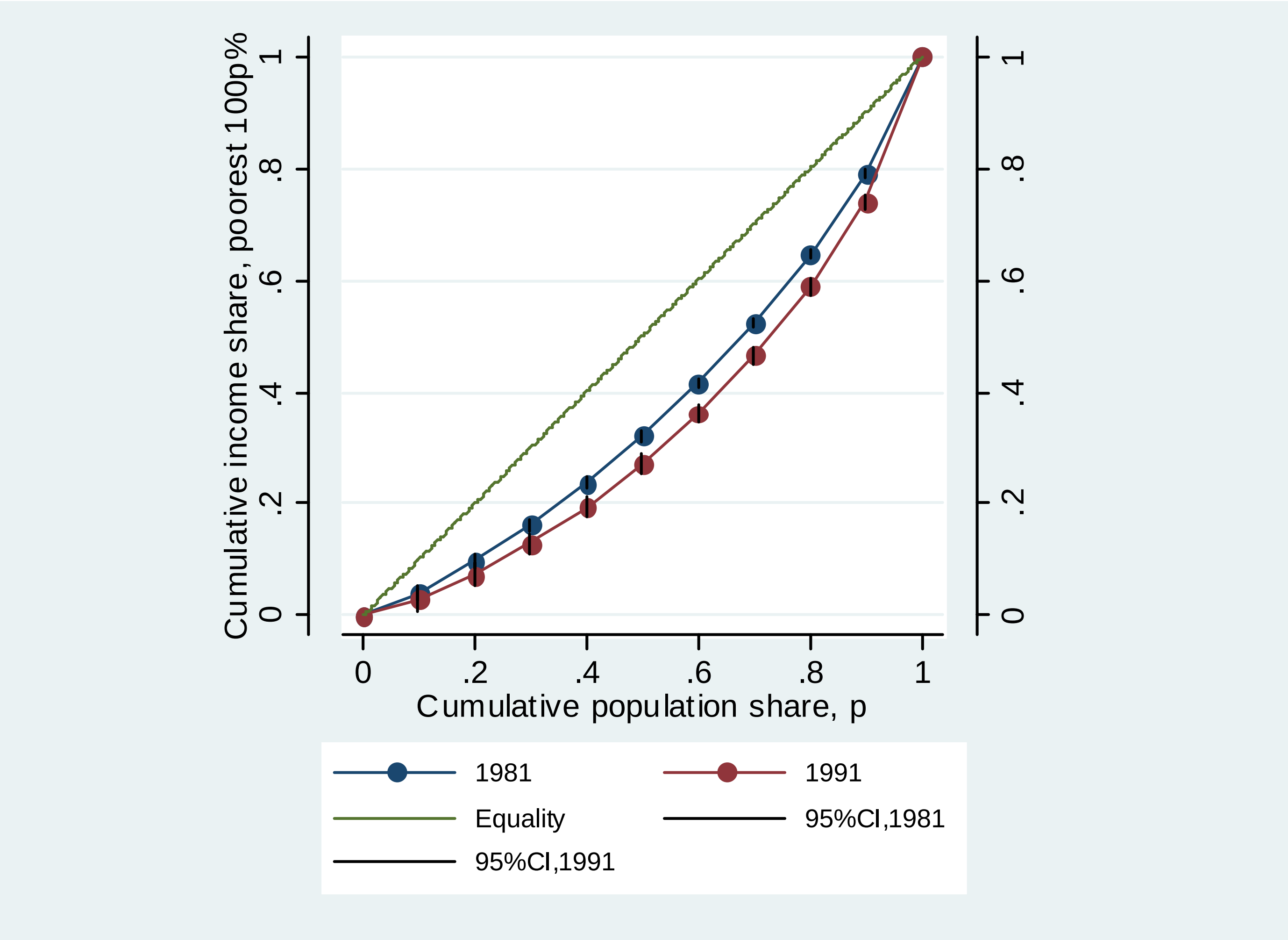
```
. svylorenz x if year == 1981, pvar(p81) lvar(rl81) selvar(se81)
. svylorenz x if year == 1991, pvar(p91) lvar(rl91) selvar(se91)

. 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)
```

Lorenz curve comparisons with CIs (2)



Note overlapping CIs at small values of *p*

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 SEs for poverty indices

```
. program define pov91, rclass
  1.          povdeco x [aw = wgt], pline($z_91)
        // version 5 program that leaves results in global macros
  2.          return scalar fgt0 = $S_FGT0
  3.          return scalar fgt1 = $S_FGT1
  4.          return scalar fgt2 = $S_FGT2
  5. end
.
. preserve

. drop if (missing(x) | year != 1991 )
(18763 observations deleted)
```

povdeco is not yet rclass, and `bootstrap` does not allow weights, so write wrapper program

You need to ensure that the bootstrap sample consists only of obs with non-missing values or excluded values on all the variables referred to in the command,
One way is to `preserve` the data, drop the values, and then `restore` the original data set

Bootstrapped SEs for poverty indices (2)

```
. bootstrap fgt0 = r(fgt0) fgt1 = r(fgt1) fgt2 = r(fgt2), ///
>          reps(250) cluster(hrn) : pov91
(running pov91 on estimation sample)
<output omitted>
Bootstrap replications (250)

Bootstrap results                          Number of obs      =      6468
                                           Number of clusters =      5254
                                           Replications       =       250

      command:  pov91
         fgt0:  r(fgt0)
         fgt1:  r(fgt1)
         fgt2:  r(fgt2)
```

The bootstrap SEs turn out to be very similar to the linearized SEs: cf. earlier estimates

```
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        fgt0 |   .2021312   .0067318    30.03   0.000     .1889371    .2153253
        fgt1 |   .0509199   .0022253    22.88   0.000     .0465584    .0552815
        fgt2 |   .0221505   .0014815    14.95   0.000     .0192469    .0250542
------------------------------------------------------------------------------
```

Bootstrapped SEs for inequality indices

1. Write wrapper program to retrieve results from `ineqdeco`
2. Drop observations not to be used in the bootstrapping

```
. prog define ineq, rclass
  1.          ineqdeco x [aw = wgt]
  2.          ret scalar gini = $S_gini
  3.          ret scalar ge0 = $S_i0
  4. end

. preserve

. drop if  (missing(x) | x <= 0 | year != 1991 )
(18783 observations deleted)
```

Bootstrapped SEs for inequality indices (2)

```
. * 250 reps
. bootstrap gini = r(gini) ge0 = r(ge0) , ///
>          reps(250) cluster(hrn) : ineq
(running ineq on estimation sample)
<output omitted>

Bootstrap replications (250)
Bootstrap results                               Number of obs      =      6448
                                                Number of clusters =      5237
                                                Replications       =       250

      command:  ineq
         gini:  r(gini)
          ge0:  r(ge0)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gini |   .3346479   .0050705    66.00   0.000     .3247099    .3445859
         ge0 |   .1952363   .0062402    31.29   0.000     .1830057    .2074669
------------------------------------------------------------------------------
```

Again, bootstrap SEs happen to be very similar to the linearized ones.

Envoi

- A fairly comprehensive suite of programs is available in Stata for many of the methods conventionally used for ‘descriptive’ analysis of distributions
- All the methods rely on you having ‘good’ data and choices from the Checklist!

What next?

- SPJ's work in progress:
  - updating programs to version 8.2 or later
- SPJ's potential future work:
  - Variance estimation using linearization methods for
    - Quantiles/CDF
    - Poverty indices with 'endogenous' poverty lines
    - Generalized Lorenz curves and TIP curves
  - Measures for income mobility and poverty dynamics
- Bigger issues:
  - multiple comparison tests and stochastic dominance checks
  - Using weights when bootstrapping
  - Etc. etc.