University of Essex

Fitting distributions by maximum likelihood

Stephen P. Jenkins

Institute for Social & Economic Research

University of Essex

Email: stephenj@essex.ac.uk

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

Outline

Fitting Generalized Beta of the Second Kind, Singh-Maddala, Dagum, and lognormal distributions to data on income or any other skewed variable of interest:

- Motivation and background
- Programs to fit the models by ML (`method lf`, v. 8.2) and predict :
  - `gb2fit`, `smfit`, `dagumfit`, `lognfit`
  - `gb2pred`, `smpred`, `dagumpred`, `lognpred`
- Assessment of goodness of fit, numerically and graphically (q-q and p-p plots using programs written by Nick Cox: `{p|q}gb2`, `{p|q}sm`, `{p|q}dagum`, `{p|q}logn`)
- Allowing parameters to vary with covariates
- Maximization issues

Motivation and background

- "Some standard functional forms claim attention, not only for their suitability in modelling some features of many empirical income distributions, but also because of their role as equilibrium distributions in economic processes." (Cowell, 2000: 145)
- Particularly useful is the GB2 family: "… the generalized beta of the second kind provided the best relative fit and included many other distributions as special or limiting cases." (McDonald, 1984: 660). See also Brachman et al.(1996), using unit record data.

GB2 distribution tree

<u>4 parameters</u>

GB2: a, b, p, q

<u>3 parameters</u>

S-M: a, b, q

Dagum: a, b, p

b: scale parameter

a, p, q: shape parameters

S-M = Burr Type 12

Dagum = Burr Type 3

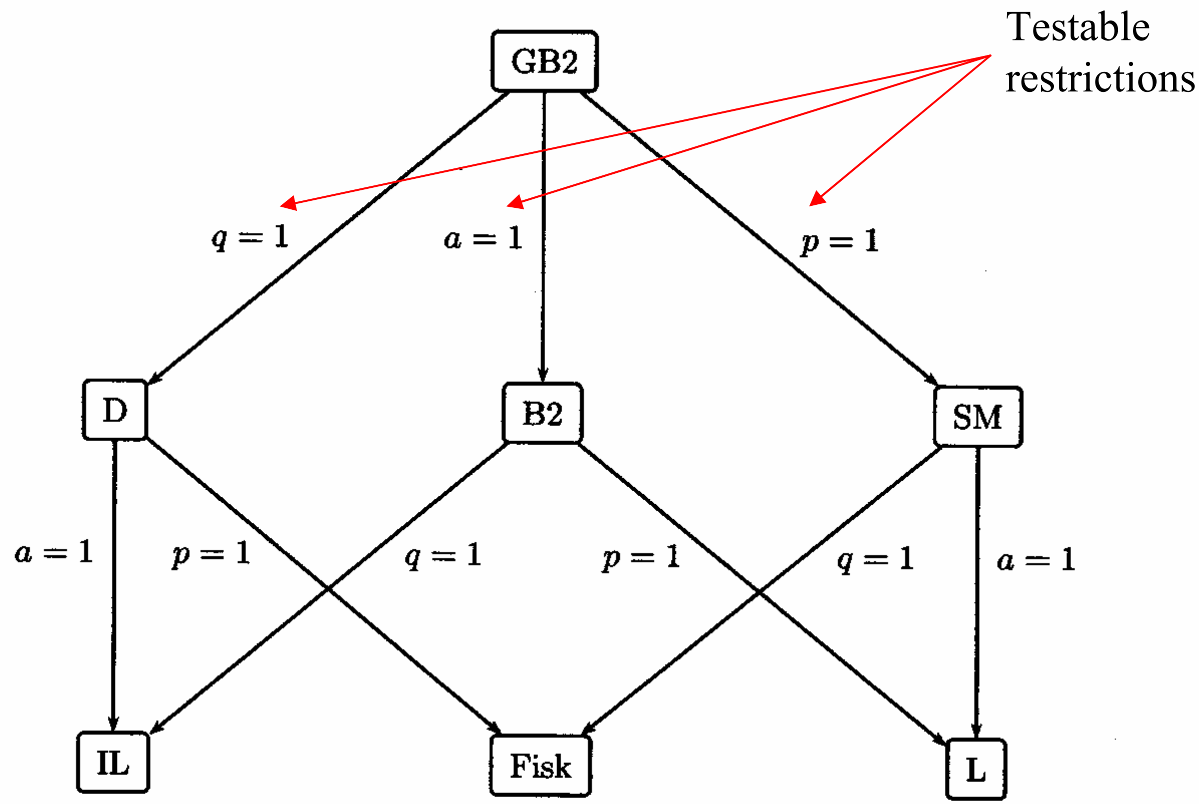


Figure 4 Beta-type size distributions and their interrelations: generalized beta distribution of the second kind (GB2), Dagum distribution (D), beta distribution of the second kind (B2), Singh–Maddala distribution (SM), inverse Lomax distribution (IL), Fisk (log-logistic) distribution (Fisk), Lomax distribution (L).

Source: Kleiber & Kotz (2003: 188)

Note: Lognormal (2 parameters) is a limiting case of GB2.

Existing programs & extensions

- `smfit`, `dagumfit` (Jenkins 1999), with `{p|q}sm`, `{p|q}dagum` (Cox 1999)
- Extensions here:
  - New distributions (GB2, lognormal)
  - Update to version 8.2 (from version 4)
  - Allow parameters to depend on covariates
  - Prediction program, and more options
  - [NJC] associated p-p and q-q plot programs
- `betafit`, `gumbelfit`, `gammafit` (Cox & Jenkins, via SSC; code based on these new programs)

Syntax: gb2fit, gb2pred

```
gb2fit var [weight] [if exp] [in range]
  [, avar(varlist1) bvar(varlist2)
  pvar(varlist3) qvar(varlist4) stats
  poorfrac(#) cdf(cdfname) pdf(pdfname)
  robust cluster(varname) svy level(#)
  maximize_options svy_options]
```

NB some not available if covariates specified

```
gb2pred [if exp] [in range] [ ,
  aval(value list1) bval(value list2)
  pval(value list3) qval(value list4)
  poorfrac(#) cdf(cdfname) pdf(pdfname) ]
```

with analogous syntax for the other programs.

`*fit` are byable, with many saved results.

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

Data used to illustrate programs

Goodman & Webb (1994) 'HBAI' subfiles derived from UK Family Expenditure Survey. 31 cross-sections 1961–1991. Focus on 1991 here.

– Data from http://www.data-archive.ac.uk

- `inc`: household post-benefit post-tax income, equivalized using McClements scale
- `wgt`: weight so data represent UK population of individuals (adults & kids)
- `but`: family ('benefit unit') type: 6 types

  couple pensioner, single pensioner, couple with child(ren),
  couple no child, single with child(ren), single no child.

- Data organised by family (may have >1 family per household, i.e. repeated obs on income within each household)

Illustration of **`gb2fit`** in action (1)

First get some non-parametric reference points for mean, s.d., inequality indices, etc. (output used later):

```
. su inc [w=wgt] if inc > 0, de
. // low income cut-off = 60% median
. local z = .6 * r(p50)
. di "Poverty line =  " `z'
. sumdist inc [w=wgt] if inc > 0, n(20)
. ineqdeco inc [w=wgt] if inc > 0
. povdeco inc [w=wgt] if inc > 0, pl(`z')
. akdensity inc [aw=wgt] if inc < 1000, n(100)
  saving(kpdf, replace) gen(y kpdf )
. lab var kpdf "Adaptive kdensity pdf"
```

```
. gb2fit inc [aw=wgt], stats poorfrac(`z')
  cluster(hrn)  pdf(gb2pdf) cdf(gb2cdf)
```

Repeated values of `inc` per household

University of Essex

```
Warning: inc has 20 values = 0; not used in calculations

initial:       log pseudo-likelihood =     -<inf>  (could not be evaluated)
feasible:      log pseudo-likelihood = -55721.411
rescale:       log pseudo-likelihood = -55721.411
rescale eq:    log pseudo-likelihood = -40305.129
Iteration 0:   log pseudo-likelihood = -40305.129
Iteration 1:   log pseudo-likelihood = -39963.522  (not concave)
Iteration 2:   log pseudo-likelihood = -39959.993  (not concave)
Iteration 3:   log pseudo-likelihood = -39945.463  (not concave)
Iteration 4:   log pseudo-likelihood = -39943.539  (not concave)
Iteration 5:   log pseudo-likelihood =   -39936.3
Iteration 6:   log pseudo-likelihood = -39931.266
Iteration 7:   log pseudo-likelihood = -39929.518
Iteration 8:   log pseudo-likelihood = -39929.254
Iteration 9:   log pseudo-likelihood = -39929.242
Iteration 10:  log pseudo-likelihood = -39929.242

ML fit of GB2 distribution                          Number of obs   =       6448
                                                    Wald chi2(0)    =          .
Log pseudo-likelihood = -39929.242                  Prob > chi2     =          .

                              (standard errors adjusted for clustering on hrn)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
a            |
       _cons |   2.720336   .2477988    10.98   0.000     2.234659    3.206013
-------------+----------------------------------------------------------------
b            |
       _cons |   206.8095   8.923641    23.18   0.000     189.3194    224.2995
-------------+----------------------------------------------------------------
p            |
       _cons |    1.04128   .1641669     6.34   0.000     .7195184    1.363041
-------------+----------------------------------------------------------------
q            |
       _cons |   1.174022   .1583051     7.42   0.000     .8637496    1.484294
------------------------------------------------------------------------------
```

Potential maximization issues: see later

Close to one: test!

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

University of Essex

poorfrac option

```
Fraction with  inc < 117.0185 =   0.18912
```

stats option

```
------------------------------------------------------------
     Quantiles       Cumulative shares of
                     total inc (Lorenz ordinates)
------------------------------------------------------------
 1%   38.55093       0.00121
 5%   69.11474       0.01080
10%   90.03578       0.02791
20%  119.92347       0.07298
25%  132.72419       0.09993
30%  144.99010       0.12954
40%  169.25560       0.19653 Mode        152.54854
50%  194.88732       0.27408 Mean        234.50523
60%  224.12633       0.36324 Std. Dev.   179.94483
70%  260.57921       0.46624
75%  283.66940       0.52418 Variance     3.24e+04
80%  312.39350       0.58760 Half CV^2     0.29440
90%  408.36262       0.73849
95%  520.38703       0.83579 p90/p10       4.53556
99%  880.55814       0.94511 p75/p25       2.13728
------------------------------------------------------------
```

NB No Gini estimate for GB2: Stata does not have ${}_3F_2$ generalized hypergeometric function

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

```
. test [p]_cons=1

 ( 1)  [p]_cons = 1

           chi2(  1) =    0.06
         Prob > chi2 =    0.8015
```

Cannot reject H0: $p = 1$ (S-M)

```
. test [p]_cons=[q]_cons=1

 ( 1)  [p]_cons - [q]_cons = 0
 ( 2)  [p]_cons = 1

           chi2(  2) =    3.09
         Prob > chi2 =    0.2131
```

Cannot reject H0: $p = q = 1$ (Fisk)

```
. constraint define 1 [p]_cons = 1

. constraint define 2 [q]_cons = 1

. gb2fit inc [aw=wgt], cluster(hrn) constraint(1)
```

Can estimate imposing constraints …

… or by estimating other models directly …

University of Essex

```
ML fit of Singh-Maddala distribution                Number of obs   =       6448
                                                    Wald chi2(0)    =          .
Log pseudo-likelihood = -39929.339                  Prob > chi2     =          .
```

Singh-Maddala

```
                              (standard errors adjusted for clustering on hrn)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
a            |
       _cons |   2.794328    .081691    34.21   0.000     2.634217     2.95444
-------------+----------------------------------------------------------------
b            |
       _cons |   206.8513   8.628886    23.97   0.000      189.939    223.7636
-------------+----------------------------------------------------------------
q            |
       _cons |   1.129722   .0848093    13.32   0.000     .9634992    1.295946
------------------------------------------------------------------------------
```

<u>GB2 estimates</u>

a: 2.72
b: 206.81
p: 1.04
q: 1.17

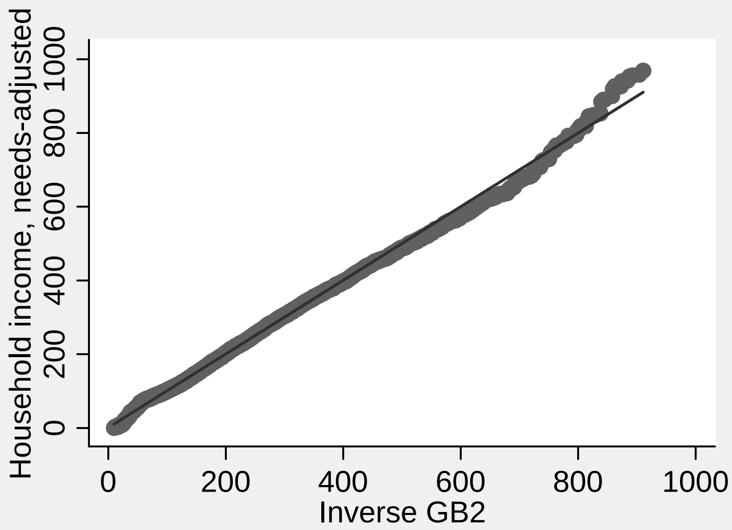
```
ML fit of Dagum distribution                        Number of obs   =       6448
                                                    Wald chi2(0)    =          .
Log pseudo-likelihood = -39930.549                  Prob > chi2     =          .
```

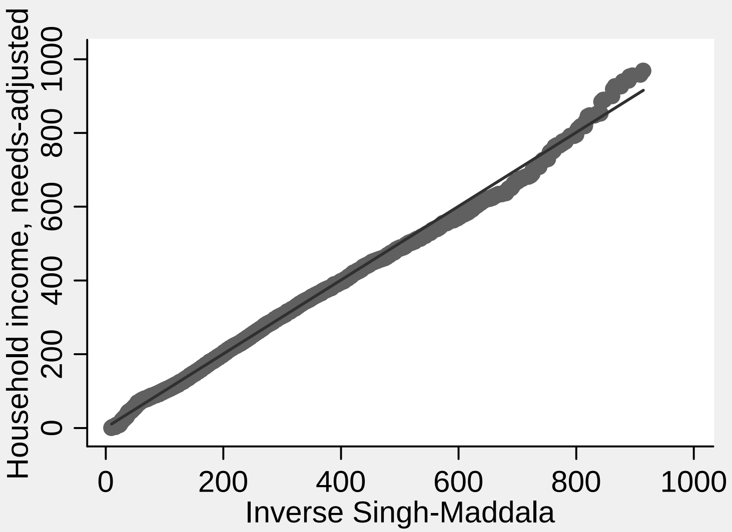
Dagum

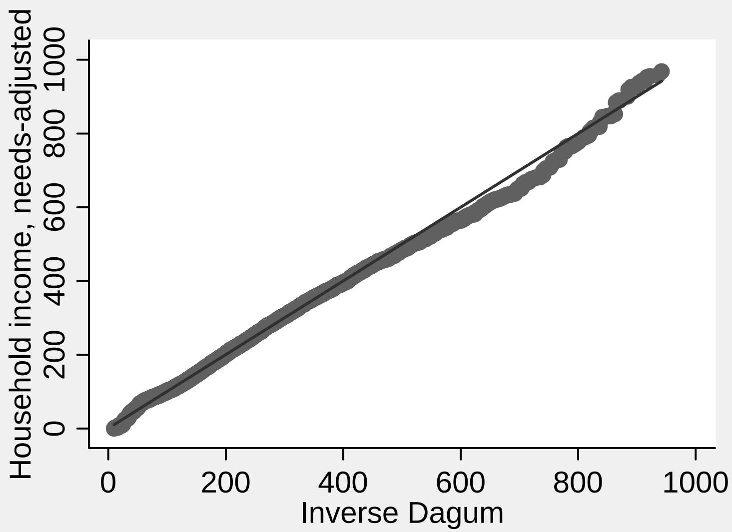
```
                              (standard errors adjusted for clustering on hrn)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
a            |
       _cons |   3.002924   .0860926    34.88   0.000     2.834185    3.171662
-------------+----------------------------------------------------------------
b            |
       _cons |   202.9363    8.55932    23.71   0.000     186.1603    219.7122
-------------+----------------------------------------------------------------
p            |
       _cons |   .9150741   .0740895    12.35   0.000     .7698614    1.060287
------------------------------------------------------------------------------
```

ISER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

Inspect fit of each model graphically with q-q plots: **`qgb2`**, **`qsm`**, **`qdagum`**







Fit generally looks good; worst at very high incomes

$p05 \approx 79$

$p50 \approx 195$

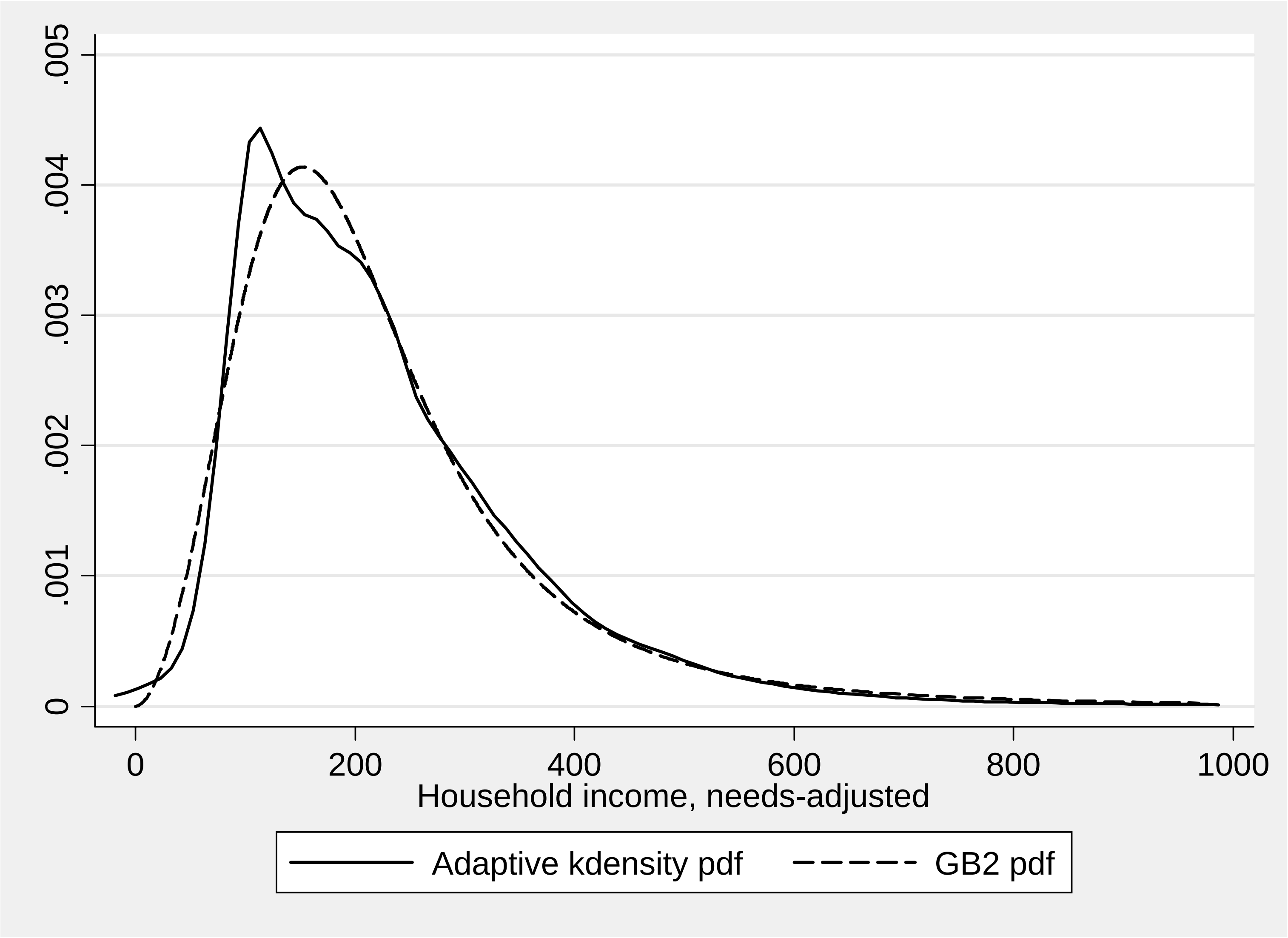
$p95 \approx 504$

$p99 \approx 942$

Could also use corresponding p-p plot programs

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

Estimates of pdf: non-parametric versus GB2



Goodness of fit depends on the eye of the beholder?

Goodness of fit again: numerically

Distributional summary statistics					
Statistic	Non-par.	GB2	S-M	Dagum	Lognormal
Mean	234.54	234.51	234.60	235.70	237.39
*p*05	79.12	69.11	68.96	69.10	66.90
*p*25	127.84	132.72	132.82	133.09	124.97
*p*50	195.03	194.88	194.92	194.68	192.94
*p*75	287.51	283.67	283.44	282.74	297.86
*p*95	504.11	520.39	520.51	524.82	556.44
% poor	20.18	18.91	18.89	18.77	21.87
Std. dev.	173.74	179.94	183.94	193.21	170.17
$\frac{1}{2}CV^2$	0.274	0.294	0.307	0.336	0.257
Gini	0.335	n.a.	0.337	0.340	0.351
*p*90/*p*10	4.336	4.536	4.534	4.528	5.209
*p*75/*p*25	2.249	2.137	2.134	2.124	2.384
*L*05	1.22	1.08	1.07	1.07	1.97
*L*25	10.14	9.99	9.99	9.96	13.81
*L*50	27.34	27.41	27.40	27.30	33.92
*L*75	52.58	52.42	52.39	52.12	60.25
*L*95	83.71	83.58	83.52	83.13	89.07

Notes. *p*XX: quantiles. *L*XX: Lorenz ordinates (%).
Poverty line = 60% median income.

Allow parameters to vary with covariates

Cf. Biewen & Jenkins (2004). Fit distributions with parameters depending on covariates to data for US, Germany & GB. Decomposition of cross-national differences in poverty rates into terms representing differences in coefficients and differences in proportions in each subgroup.

Simple example here: S-M model parameters differing by family type

```
. ta but [aw=wgt], ge(ftype) // create DVs for 6 family types
```

ftype1 = 1: couple pensioner (reference category in regression)

ftype2 = 1: single pensioner

ftype3 = 1: couple with child(ren)

ftype4 = 1: couple no child

ftype5 = 1: single with child(ren)

ftype6 = 1: single no child

```
. smfit inc [aw=wgt], cluster(hrn) a(ftype2-ftype6) b(ftype2-
  ftype6) q(ftype2-ftype6) difficult
. <iteration log omitted>
```

University of Essex

ML fit of Singh-Maddala distribution

Number of obs = 6448

Wald chi2(5) = 60.57

Log pseudo-likelihood = -39322.92

Prob > chi2 = 0.0000

(standard errors adjusted for clustering on hrn)

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
a						
ftype2	-.0141554	1.262981	-0.01	0.991	-2.489552	2.461241
ftype3	-3.841888	1.083919	-3.54	0.000	-5.966331	-1.717446
ftype4	-3.864617	1.080507	-3.58	0.000	-5.982371	-1.746863
ftype5	.2971847	1.732682	0.17	0.864	-3.09881	3.69318
ftype6	-4.072256	1.07808	-3.78	0.000	-6.185255	-1.959257
_cons	6.617466	1.073933	6.16	0.000	4.512596	8.722336
b						
ftype2	-.8024619	6.341398	-0.13	0.899	-13.23137	11.62645
ftype3	116.8537	16.74765	6.98	0.000	84.0289	149.6785
ftype4	213.9446	21.69987	9.86	0.000	171.4136	256.4755
ftype5	-5.582166	7.379927	-0.76	0.449	-20.04656	8.882225
ftype6	178.2679	22.48037	7.93	0.000	134.2072	222.3287
_cons	104.3952	5.558	18.78	0.000	93.5017	115.2887
q						
ftype2	.0498175	.0812713	0.61	0.540	-.1094714	.2091064
ftype3	1.091651	.2017097	5.41	0.000	.6963077	1.486995
ftype4	1.147459	.1960433	5.85	0.000	.7632217	1.531697
ftype5	.1034508	.121696	0.85	0.395	-.135069	.3419705
ftype6	1.415773	.2398489	5.90	0.000	.9456777	1.885868
_cons	.2952924	.0662383	4.46	0.000	.1654677	.425117

Here are the estimates,

but what do they imply about the distributions for different groups?

ISER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

```
. // parameter values for each obs in the sample
. matrix a = e(b_a)
. matrix b = e(b_b)
. matrix q = e(b_q)
. matrix score a_i = a, eq(a)
. matrix score b_i = b, eq(b)
. matrix score q_i = q, eq(q)
. list hrn but a_i b_i q_i in 1/15, noobs
```

Model parameters, conditional on characteristics

Manually (as here) or, use prediction program (coming next)

hrn	but	a_i	b_i	q_i
914018	couple w	2.775578	221.2489	1.386944
910414	single p	6.603311	103.5927	.3451098
914670	couple w	2.775578	221.2489	1.386944
911383	couple w	2.775578	221.2489	1.386944
910230	couple w	2.775578	221.2489	1.386944
911140	single n	2.545211	282.6631	1.711065
914129	couple n	2.752849	318.3398	1.442752
915231	couple w	2.775578	221.2489	1.386944
910156	couple n	2.752849	318.3398	1.442752
910085	couple w	2.775578	221.2489	1.386944
917575	couple n	2.752849	318.3398	1.442752
911669	couple n	2.752849	318.3398	1.442752
915201	couple w	2.775578	221.2489	1.386944
914036	single w	6.914651	98.81301	.3987432
914443	couple p	6.617466	104.3952	.2952923

University of Essex

Statistics for given covariate values: **gb2pred, smpred, dagumpred, lognpred**

```
. // fill in values for: ftype2, ftype3, ftype4, ftype5, ftype6, constant
. smpred, a(0,0,0,0,0,1) b(0,0,0,0,0,1) q(0,0,0,0,0,1) poorfrac(`z') // couple pensioner
------------------------------------------------------------
     Quantiles        Cumulative shares of
                      total inc (Lorenz ordinates)
------------------------------------------------------------
 1%    62.79896       0.00262
 5%    81.20567       0.01686
10%    91.85463       0.03779
20%   106.32747       0.08570
25%   112.59280       0.11207
30%   118.75516       0.13993
40%   131.64446       0.20019 Mode         115.04276
50%   146.60053       0.26710 Mean         207.58730
60%   165.69503       0.34209 Std. Dev.    649.67642
70%   192.81477       0.42800
75%   211.92418       0.47664 Variance      4.22e+05
80%   237.73570       0.53062 Half CV^2      4.89737
90%   339.15746       0.66536 Gini coeff.    0.37439
95%   483.58990       0.76144 p90/p10        3.69233
99%    1.10e+03       0.89127 p75/p25        1.88222
------------------------------------------------------------

Fraction with inc < 117.0185 =    0.28594
```

You can use `*pred` repeatedly to derive predictions for different covariate combinations (i.e. different family types here)

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

The svy option (after svyset)

- Revisiting the earlier example, note that

```
. gb2fit inc [aw=wgt], cluster(hrn)
```

produces the same estimates, given structure of this data set, as:

```
. svyset [pw=wgt], psu(hrn)
. gb2fit inc, svy
```

- Example from dataset with complex survey design (British Household Panel Survey, wave 1). Observations on equivalised household income for individuals (multiple individuals per household)

```
. svyset [pw=xewght], psu(psu) strata(strata)
pweight is xewght
strata is strata
psu is psu

. gb2fit net, svy
```

svy option in action: BHPS illustration

```
ML fit of GB2 distribution

pweight:  xewght                                  Number of obs    =      11616
Strata:   strata                                  Number of strata =         75
PSU:      psu                                     Number of PSUs   =        250
                                                  Population size  = 11542.016
                                                  F(   0,     176) =          .
                                                  Prob > F         =          .

------------------------------------------------------------------------------
         net |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
a            |
       _cons |   1.861342   .2113467     8.81   0.000     1.444225    2.278458
-------------+----------------------------------------------------------------
b            |
       _cons |    234.628    22.2282    10.56   0.000     190.7582    278.4979
-------------+----------------------------------------------------------------
p            |
       _cons |   1.395198   .2484427     5.62   0.000     .9048688    1.885528
-------------+----------------------------------------------------------------
q            |
       _cons |   2.954858   .6620281     4.46   0.000     1.648271    4.261445
------------------------------------------------------------------------------
```


Second-level clustering within households is automatically accounted for

Maximization issues

- GB2 likelihood sometimes difficult to maximize
  - E.g. estimation of model without covariates for each of 31 years 1961–1991 gave 6 years in which default NR algorithm failed to reach maximum (non-concavities)
  - Maximum reached after fiddling with options:

1963: `tech(nr 1 dfp 1)` (estimated variance was negative!)
1964: `tech(nr 1 dfp 20)` (estimated variance was negative!)
1965: `tech(bfgs 1 nr 5) difficult`
1988: `tech(nr 1 dfp 1)`
1989: `tech(nr 1 dfp 1) difficult`
1990: `tech(nr 1 dfp 1)`

- Would help to impose non-negativity constraints on each parameter? (But what if covariates used?)
- Starting values other than those chosen by `ml search`? (Which?)

Concluding remarks

- Suite of programs, which could be easily extended to include other distributions
- Can calculate additional post-estimation statistics (e.g. SEs for distributional summary statistics using delta method)
- Illustrates value of the new features added to `ml` from version 8.1 onwards
  - Learn them, as I did, using Gould et al. (2003)!
- Aiming to write up for submission to *SJ*, and to release programs via that medium

University of Essex

References

Biewen, M. and Jenkins, S.P. (2004). Accounting for differences in poverty between the USA, Britain and Germany. *Empirical Economics*, forthcoming. Revised version of ISER WP 2002-14. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2002-14.pdf

Brachman, K. Stich, A., and Trede, M. (1996). Evaluating parametric income distribution models. *Allegemeine Statistiches Archiv* 80: 285–298.

Cowell, F.A. (2000). Measurement of inequality. In: A.B. Atkinson & F. Bourguignon (eds), *Handbook of Income Distribution*. Amsterdam: Elsevier.

Cox, N.J. (1999). Diagnostic plots for Singh-Maddala and Dagum distributions fitted by MLE. *Stata Technical Bulletin* STB-48, 2–4.

Goodman, A. and Webb, S. (1994). *For Richer, For Poorer. The Changing Distribution of Income in the United Kingdom, 1961-91*, Commentary No. 42. London: Institute for Fiscal Studies. Abridged version in: *Fiscal Studies* 15: 29–62.

Gould, W., Pitblado, J., and Sribney, W. (2003). *Maximum Likelihood Estimation with Stata*, second edition. College Station, TX: Stata Press.

Jenkins, S.P. (1999). Fitting Singh-Maddala and Dagum distributions by maximum likelihood. *Stata Technical Bulletin* STB-48, 19–25.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken, NJ: John Wiley.

McDonald, J.B. (1984). Some generalized functions for the size distribution of income. *Econometrica* 52: 647–663.

ISER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH