Influence functions at work

Philippe Van Kerm

Luxembourg Institute of Socio-Economic Research
philippe.vankerm@liser.lu

2015 London Stata Users Group meeting

September 10–11 2015, Cass Business School, London

Introduction

- Illustration of practical uses of ‘influence function’ estimators with Stata
  1. Study structure of summary statistics (identification of ‘influential observations’)
  2. Variance estimation
  3. ‘RIF regression’
- Application to income distribution analysis

Definition

Let $\upsilon(F)$ be a statistic of interest (a functional) calculated in distribution F (the mean, a percentile, the Gini coefficient of inequality, etc.)

The *influence function* of υ is a function of y and F and is defined as (Hampel, 1974)

$$\mathrm{IF}(y; \upsilon, F) = \lim_{\epsilon \downarrow 0} \frac{\upsilon((1-\epsilon)F + \epsilon\Delta_y) - \upsilon(F)}{\epsilon}$$

The IF captures the effect on $\upsilon(F)$ of an infinitesimal 'contamination' of F at point mass y.

(Note the expected value of the IF is zero: $\int IF(y, \upsilon(F))dF(y) = 0$)

Definition (ctd.)

Expressions for $\mathrm{IF}(y; \upsilon, F)$ exist (or can be derived) for a wide range of statistics υ[1]:

... simple (linear) statistics, e.g., the mean

$$\mathrm{IF}(y; \mu, F) = y - \mu(F)$$

... and more complex (non linear) statistics, e.g., a quantile

$$\mathrm{IF}(y; Q_\theta, F) = \frac{-1}{f(Q_\theta(F))}(I(y \leq Q_\theta(F)) - \theta)$$

[1]See e.g., Essama-Nssah and Lambert (2012) for a catalogue of IFs relevant to income distribution analysis

Practical use 1

Practical use 1:

- visualising the ‘structure’ of a (possibly complex) index
- comparison of indices (think of the many inequality measures!)
- identification of influential observations (and robustness of the index)

Income inequality indicators: the Atkinson index

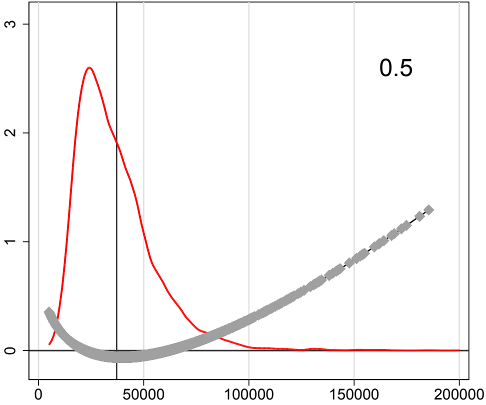
The Atkinson inequality index (Atkinson, 1970):

$$A(\epsilon) = 1 - \frac{1}{\mu}\left(\frac{1}{N}\sum_{i=1}^{N} y_i^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}}$$

for $\epsilon \geq 0$

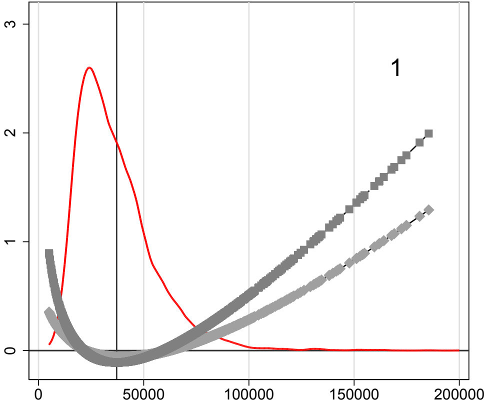
The higher ϵ, the higher 'inequality aversion'... Can we visualise that?

Income inequality indicators: the Atkinson index IF



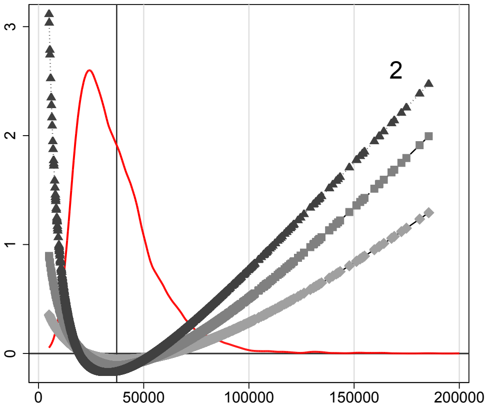
(annual household income data for Luxembourg 2012)

Income inequality indicators: the Atkinson index IF



(annual household income data for Luxembourg 2012)

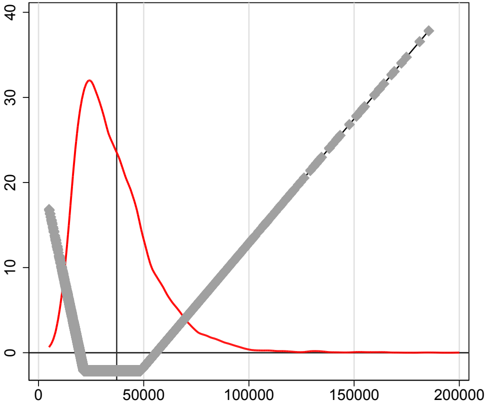
Income inequality indicators: the Atkinson index IF



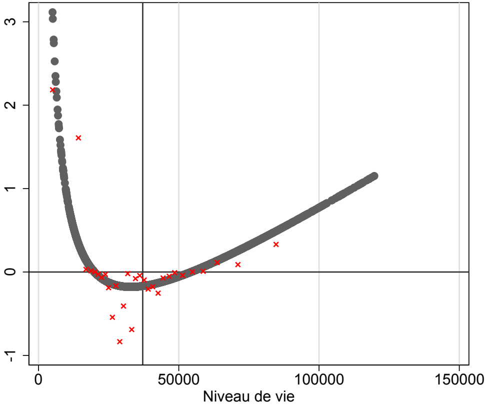
(annual household income data for Luxembourg 2012)

Income inequality: the Quintile Share Ratio

How does it compare with the Quintile (Group) Share Ratio?



An aside: comparison with delete-one jackknife influence measure



Not the same when observations are *weighted*!

Practical use 2

Practical use 2:

- estimation of the sampling variance of the index
- asymptotic approximation that works with complex non-linear statistics
- valid with complex survey design!
- (... it is all in the Stata manuals already)

Variance estimation

An asymptotic approximation of the variance of υ is given by (Hampel, 1974)

$$V(\upsilon, F) \approx \int \mathrm{IF}(y; \upsilon, F)^2 dF(y)$$

Practically boils down to estimation of a total (Deville, 1999):

$$V(\hat{\upsilon}, F) \approx V\left(\sum_{i=1}^{N} w_i \mathrm{IF}(y_i; \upsilon, \hat{F})\right)$$

... and formula well-known for the variance of a total even with complex survey design: implemented in Stata!

Variance estimation

Code template

```
svyset ...
generate rif= ...  // note:  point estimate is added to
IF
svy:  mean rif
```

Silly example with the mean:

```
svyset [pw=W] , ...
su y [aw=W]
gen rifmean = r(mean) + (y - r(mean))
svy:  mean rifmean
svy:  mean y // yes, it works!
```

Variance estimation

Built in some user-written commands

```
.         svy : inequaly nivie  , atkinson(0.5 1 2)  s80s20
(running inequaly on estimation sample)

Survey data analysis

Number of strata   =          4                Number of obs      =      5,988
Number of PSUs     =      5,988                Population size    = 515,407.14
                                               Design df          =      5,984
                                               F(   0,   5984)    =          .
                                               Prob > F           =          .

------------------------------------------------------------------------------
             |             Linearized
       nivie |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
atkp5        |
       _cons |   .0621767   .0039672    15.67   0.000     .0543995     .069954
-------------+----------------------------------------------------------------
atk1         |
       _cons |   .1174055   .0061389    19.12   0.000     .1053711      .12944
-------------+----------------------------------------------------------------
atk2         |
       _cons |   .2146223   .0083138    25.82   0.000     .1983242    .2309204
-------------+----------------------------------------------------------------
s80s20       |
       _cons |   3.953448   .1226658    32.23   0.000     3.712978    4.193917
------------------------------------------------------------------------------
```

Practical use 3

Practical use 3:

- ‘Recentered IF regression’ (Firpo et al., 2007, 2009)
- evaluate impact of covariates on distribution statistics

RIF regression

The effect of interest

For example, how do foreign households affect $\upsilon(F)$?

$$F(y) = \sum_{x \in \Omega_X} s_x \, F_x(y)$$

Consider an infinitesimal variation: swap native for foreign workers

$$G_r^{F,t,k}(y) = (s_k + t) \, F_k(y) + (s_r - t) \, F_r(y) + \sum_{x \in \Omega_X \setminus \{k,r\}} s_x \, F_x(y).$$

(Choe and Van Kerm, 2014)

What is the impact of this swap on the statistic of interest?

Methods (ctd.)

Recentered influence function estimator

Firpo et al. (2009) show that effect of interest is given by:

$$\mathrm{E}[\mathrm{RIF}(y; \upsilon, F)|X = k] - \mathrm{E}[\mathrm{RIF}(y; \upsilon, F)|X = r]$$

where $\mathrm{RIF}(y; \upsilon, F) = \upsilon(F) + \mathrm{IF}(y; \upsilon, F)$

Regression-based estimator, β in :

$$\mathrm{E}[\mathrm{RIF}(y; \upsilon, F)|X = x] = \alpha + x\beta$$

(Note: N. Fortin provides the Stata package `rifreg` for regressions on quantile, variance and Gini functionals (`http://faculty.arts.ubc.ca/nfortin/datahead.html`).)

Interpretation of RIF regression coefficients

- The RIF at y gives the influence on $\upsilon(F)$ of an infinitesimal increase in the density of the data at y
- Regression coefficients reveal how much the average influence of observations vary with X (holding other covariates constant)
- It also reveals how much $\upsilon(F)$ would respond to a change in the distribution of X in the population holding distribution of other covariates constant
  - linear approximation valid only for *marginal* changes in X!

Illustrative example 1

Data and problem

- Panel Study Liewen zu Letzebuerg 2011 (official source for poverty and inequality statistics in Luxembourg)
- Permanent panel (started in 2003) gradually converted into rotating panel from 2010:
  - new samples ('rotation groups') now added for four years (so dataset for 2010, 2011, 2012 is mix of old and new samples)
  - Concern about how much 'new samples' differ from 'old sample' (due, e.g., to attrition (and slightly different sampling frame))
  - Impact on trends in inequality and poverty indicators? Break in series when 'old sample' abandoned?

$\Longrightarrow$ Use RIF regression to check if rotation group impacts on inequality

Illustrative example 1

Code

```
svy:  inequaly nivie , atkinson(0.5 1 2)
predict rif* , rif // predict after -inequaly- gives (R)IF
svy:  regress rif1 ib9.(rot)
svy:  regress rif2 ib9.(rot)
svy:  regress rif3 ib9.(rot)
svy:  inequaly nivie , s80s20
predict rifs80s20 , rif
svy:  regress rifs80s20 ib9.(rot)
svy:  newpoverty nivie , fracmedian(.6)
predict rifh , rif
svy:  regress rifh ib9.(rot)
```

Results: Atkinson(0.5)

```
.         svy: regress rif1 ib9.(rot)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         4                  Number of obs     =      5,988
Number of PSUs     =     5,988                  Population size   = 515,407.14
                                                Design df         =      5,984
                                                F(   3,    5982)  =       2.96
                                                Prob > F          =     0.0312
                                                R-squared         =     0.0023

------------------------------------------------------------------------------
             |             Linearized
        rif1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
          1  |   .0157444   .0186107     0.85   0.398    -.0207392     .052228
          2  |  -.0048659   .0043721    -1.11   0.266    -.0134369     .003705
          3  |   .0111881   .0059868     1.87   0.062    -.0005481    .0229243
             |
       _cons |   .0575857   .0033763    17.06   0.000     .0509669    .0642044
------------------------------------------------------------------------------
```

Results: Atkinson(2)

```
.         svy: regress rif3 ib9.(rot)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =          4              Number of obs     =      5,988
Number of PSUs     =      5,988              Population size   = 515,407.14
                                             Design df         =      5,984
                                             F(   3,    5982)  =       2.04
                                             Prob > F          =     0.1056
                                             R-squared         =     0.0023

------------------------------------------------------------------------------
             |             Linearized
        rif3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
          1  |   .0246292   .0347917     0.71   0.479    -.0435751    .0928336
          2  |   -.006993   .0135387    -0.52   0.606    -.0335338    .0195478
          3  |   .0339101   .0171911     1.97   0.049     .0002094    .0676108
             |
       _cons |   .2036115    .009367    21.74   0.000     .1852488    .2219741
------------------------------------------------------------------------------
```

Results: S80/S20

```
.          svy: regress rifs80s20 ib9.(rot)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =          4                 Number of obs     =      5,988
Number of PSUs     =      5,988                 Population size   = 515,407.14
                                                Design df         =      5,984
                                                F(   3,    5982)  =       2.44
                                                Prob > F          =     0.0622
                                                R-squared         =     0.0027

------------------------------------------------------------------------------
             |             Linearized
   rifs80s20 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
           1 |   .4599947   .5221966     0.88   0.378    -.5636989    1.483688
           2 |  -.0630366   .2020467    -0.31   0.755    -.4591209    .3330477
           3 |   .5306184   .2375919     2.23   0.026     .0648526    .9963842
             |
       _cons |   3.753404   .1405622    26.70   0.000     3.477851    4.028956
------------------------------------------------------------------------------
```

Results: Poverty rate

```
.         svy: regress rifh ib9.(rot)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =          4                 Number of obs     =      5,988
Number of PSUs     =      5,988                 Population size   = 515,407.14
                                                Design df         =      5,984
                                                F(   3,    5982)  =       1.32
                                                Prob > F          =     0.2659
                                                R-squared         =     0.0020

------------------------------------------------------------------------------
             |             Linearized
        rifh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
          1  |   .0146267   .0233548     0.63   0.531    -.0311571    .0604105
          2  |   .0069663   .0211085     0.33   0.741     -.034414    .0483466
          3  |    .043081   .0221201     1.95   0.052    -.0002823    .0864443
             |
       _cons |   .1342025   .0129755    10.34   0.000     .1087658    .1596393
------------------------------------------------------------------------------
```

Illustrative example 2

Effect of foreign households on inequality and poverty?

- Effect of a marginal increase in share of foreign-headed households on indicators
  - assuming no change in income structure otherwise
- Condition on age of foreign households

Results: Atkinson(0.5)

```
.         svy: regress rif1 ib9.(rot) i.(chme11)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =          4                  Number of obs      =      5,987
Number of PSUs     =      5,987                  Population size    = 515,119.97
                                                 Design df          =      5,983
                                                 F(   6,    5978)   =       7.42
                                                 Prob > F           =     0.0000
                                                 R-squared          =     0.0099

------------------------------------------------------------------------------
             |             Linearized
        rif1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
          1  |   .0142397   .0183501     0.78   0.438    -.0217331    .0502126
          2  |  -.0062364   .0043078    -1.45   0.148    -.0146812    .0022083
          3  |   .0080582   .0059128     1.36   0.173     -.003533    .0196495
             |
      chme11 |
   Portugais |   .0135398   .0039052     3.47   0.001     .0058841    .0211955
Autres UE-15 |   .0262902   .0178432     1.47   0.141     -.008689    .0612694
   Non UE-15 |   .0418941   .0086454     4.85   0.000     .0249461    .0588421
             |
       _cons |   .0476424   .0050467     9.44   0.000     .0377491    .0575358
------------------------------------------------------------------------------
```

Results: Atkinson(0.5)

```
.         svy: regress rif1 ib9.(rot) i.(chme11) ib6.(chme09)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         4                  Number of obs     =      5,987
Number of PSUs     =     5,987                  Population size   = 515,119.97
                                                Design df         =      5,983
                                                F(  10,    5974)  =       5.90
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0108

------------------------------------------------------------------------------
             |             Linearized
        rif1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
          1  |   .0141385   .0178513     0.79   0.428    -.0208565    .0491335
          2  |  -.0061202   .0043087    -1.42   0.156    -.0145667    .0023264
          3  |   .0082485   .0059444     1.39   0.165    -.0034047    .0199016
             |
      chme11 |
   Portugais |   .0159073   .0039841     3.99   0.000      .008097    .0237177
Autres UE-15 |   .0269482   .0181449     1.49   0.138    -.0086224    .0625188
   Non UE-15 |   .0442646   .0085935     5.15   0.000     .0274182     .061111
             |
      chme09 |
     [16-24] |   .0032088   .0138918     0.23   0.817    -.0240242    .0304418
     [25-34] |  -.0125041   .0078015    -1.60   0.109    -.0277978    .0027897
     [35-49] |  -.0050532    .007715    -0.65   0.512    -.0201773    .0100708
     [50-64] |   .0021521   .0129697     0.17   0.868    -.0232733    .0275774
             |
       _cons |   .0499364   .0094564     5.28   0.000     .0313985    .0684744
------------------------------------------------------------------------------
```

Results: Atkinson(2)

```
.         svy: regress rif3 ib9.(rot) i.(chme11) ib6.(chme09)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         4                Number of obs     =      5,987
Number of PSUs     =     5,987                Population size   = 515,119.97
                                              Design df         =      5,983
                                              F(  10,    5974)  =       2.68
                                              Prob > F          =     0.0029
                                              R-squared         =     0.0173

------------------------------------------------------------------------------
             |             Linearized
        rif3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
          1  |   .0197437    .033453     0.59   0.555    -.0458363    .0853236
          2  |  -.0112723   .0133852    -0.84   0.400    -.0375122    .0149676
          3  |   .0245481   .0159634     1.54   0.124    -.0067459    .0558422
             |
      chme11 |
   Portugais |   .0076757   .0114413     0.67   0.502    -.0147533    .0301047
Autres UE-15 |   .0634587   .0338566     1.87   0.061    -.0029124    .1298298
   Non UE-15 |   .1332594   .0373124     3.57   0.000     .0601136    .2064052
             |
      chme09 |
     [16-24] |    .046107   .0539939     0.85   0.393    -.0597405    .1519544
     [25-34] |  -.0031152   .0206771    -0.15   0.880    -.0436499    .0374194
     [35-49] |   .0021401   .0159684     0.13   0.893    -.0291637    .0334439
     [50-64] |   .0165315   .0241836     0.68   0.494    -.0308771      .06394
             |
       _cons |   .1764981   .0180461     9.78   0.000     .1411213    .2118749
------------------------------------------------------------------------------
```

Results: Poverty rate

```
.         svy: regress rifh ib9.(rot) i.(chme11) ib6.(chme09)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =          4                Number of obs     =      5,987
Number of PSUs     =      5,987                Population size   = 515,119.97
                                               Design df         =      5,983
                                               F(  10,    5974)  =       2.89
                                               Prob > F          =     0.0013
                                               R-squared         =     0.0142

------------------------------------------------------------------------------
             |             Linearized
        rifh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
           1 |   .0114206   .0234601     0.49   0.626    -.0345697    .0574109
           2 |   .0040279   .0209573     0.19   0.848     -.037056    .0451117
           3 |   .0380812   .0219422     1.74   0.083    -.0049334    .0810959
             |
      chme11 |
   Portugais |  -.0770121   .0238951    -3.22   0.001    -.1238552    -.030169
Autres UE-15 |   .0329958    .019137     1.72   0.085    -.0045195    .0705112
  Non UE-15  |   .0465431   .0511668     0.91   0.363    -.0537624    .1468486
             |
      chme09 |
     [16-24] |   .0411687   .0877057     0.47   0.639     -.130766    .2131034
     [25-34] |   .0480566   .0257379     1.87   0.062     -.002399    .0985123
     [35-49] |   .0379237   .0192831     1.97   0.049     .0001218    .0757256
     [50-64] |   .0562376   .0190103     2.96   0.003     .0189707    .0935046
             |
       _cons |   .1004148   .0159395     6.30   0.000     .0691676     .131662
------------------------------------------------------------------------------
```

Results: Poverty rate (fixed poverty line)

```
.         svy: regress rifhbis1 ib9.(rot) i.(chme11) ib6.(chme09)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         4                  Number of obs     =      5,987
Number of PSUs     =     5,987                  Population size   = 515,119.97
                                                Design df         =      5,983
                                                F(  10,    5974)  =      16.31
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1074

------------------------------------------------------------------------------
             |             Linearized
    rifhbis1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rot |
          1  |   .0178995   .0213374     0.84   0.402    -.0239295    .0597284
          2  |   .0059625   .0192948     0.31   0.757    -.0318624    .0437873
          3  |    .038296   .0218492     1.75   0.080    -.0045364    .0811283
             |
      chme11 |
   Portugais |   .2040203   .0231579     8.81   0.000     .1586225    .2494181
Autres UE-15 |    .052723   .0189874     2.78   0.006     .0155009    .0899452
  Non UE-15  |   .3275548   .0547087     5.99   0.000     .2203059    .4348036
             |
      chme09 |
    [16-24]  |   .1271475   .0982018     1.29   0.195    -.0653634    .3196585
    [25-34]  |    .047582   .0231179     2.06   0.040     .0022626    .0929015
    [35-49]  |   .0561238    .015962     3.52   0.000     .0248325    .0874151
    [50-64]  |   .0382346    .014837     2.58   0.010     .0091487    .067320 4
             |
       _cons |   .0181208   .0122119     1.48   0.138    -.0058189    .042060 6
------------------------------------------------------------------------------
```

Atkinson, A. B. (1970), 'On the measurement of inequality', *Journal of Economic Theory* **2**(3), 244–263.

Choe, C. and Van Kerm, P. (2014), Foreign workers and the wage distribution: Where do they fit in?, CEPS/INSTEAD Working Paper 2014-02, Luxembourg Institute of Socio-Economic Research, Esch-sur-Alzette, Luxembourg.

Deville, J.-C. (1999), 'Variance estimation for complex statistics and estimators: linearization and residual techniques', *Survey Methodology* **25**, 193–204.

Essama-Nssah, B. and Lambert, P. J. (2012), Influence functions for policy impact analysis, *in* J. A. Bishop and R. Salas, eds, 'Inequality, Mobility and Segregation: Essays in Honor of Jacques Silber', Vol. 20 of *Research on Economic Inequality*, Emerald Group Publishing, chapter 6, pp. 135–159.

Firpo, S., Fortin, N. M. and Lemieux, T. (2007), Unconditional quantile regressions, Technical Working Paper 339, National Bureau of Economic Research, Cambridge MA, USA.

Firpo, S., Fortin, N. M. and Lemieux, T. (2009), 'Unconditional quantile regressions', *Econometrica* **77**(3), 953–973.

Hampel, F. R. (1974), 'The influence curve and its role in robust estimation', *Journal of the American Statistical Association* **69**(346), pp. 383–393.

Support from the Luxembourg ‘Fonds National de la Recherche’ is gratefully acknowledged (project ‘Tax-benefit systems, employment structures and cross-country differences in income inequality in Europe: a micro-simulation approach–SIMDECO’).