Influence functions at work

Philippe Van Kerm

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

2016 German Stata Users Group meeting

June 10 2016, GESIS, Cologne

LISER
LUXEMBOURG INSTITUTE OF SOCIO-ECONOMIC RESEARCH

Introduction

- Illustration of practical uses of ‘influence function’ estimators with Stata
  1. Study *structure* of some summary statistics of interest—e.g., ‘social indicators’
     - identification of ‘influential observations’
     - robustness properties
  2. *Variance* estimation and testing
  3. “RIF *regression*”
- Application to income distribution analysis

Definition

Let $\upsilon(F)$ be a statistic of interest (a functional) calculated in distribution F, e.g. the mean, the median, a percentile, the Gini coefficient of inequality, a 'top (income) share', a correlation or regression coefficient, 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.

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

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

[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):

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

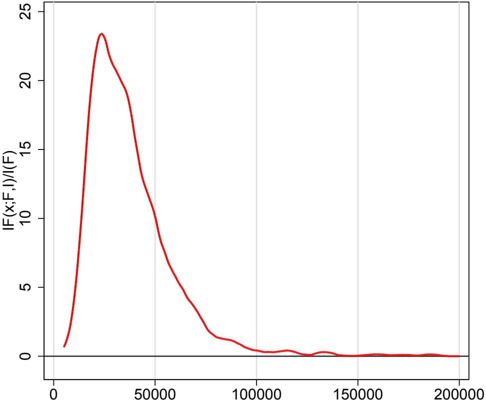
for $\epsilon \geq 0$

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

$$IF(y; A(\epsilon)) = \mu_\epsilon^{\frac{\epsilon}{1-\epsilon}} \frac{(y^{1-\epsilon} - \mu_\epsilon)}{(\epsilon - 1)\mu} + \mu_\epsilon^{\frac{1}{1-\epsilon}} \frac{(y-\mu)}{\mu^2}$$

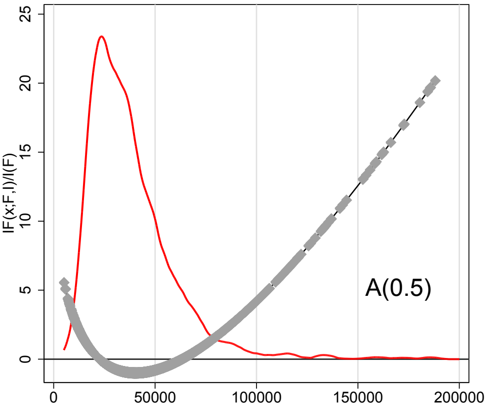
(Cowell and Flachaire, 2007)

Income inequality indicators: the Atkinson index IF



(annual household income data for Luxembourg 2012)

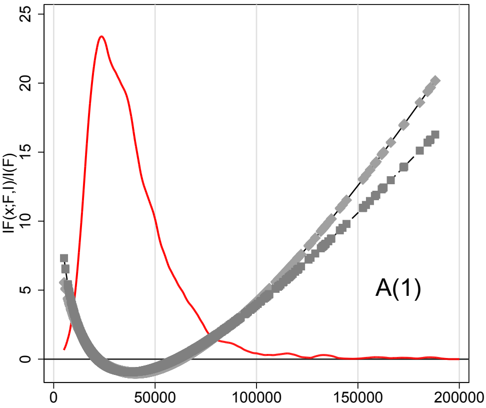
Income inequality indicators: the Atkinson index IF



(annual household income data for Luxembourg 2012)

Income inequality indicators: the Atkinson index IF

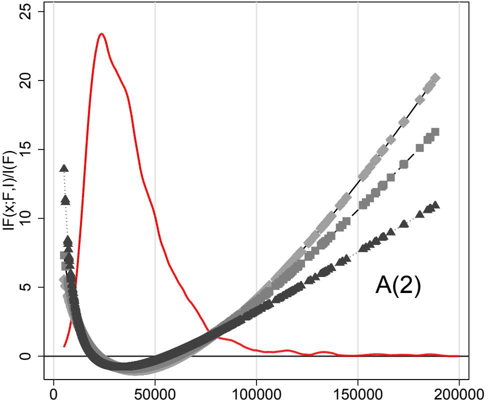
Changing sensitivity—‘inequality aversion parameters’



(annual household income data for Luxembourg 2012)

Income inequality indicators: the Atkinson index IF

Changing sensitivity—'inequality aversion parameters'

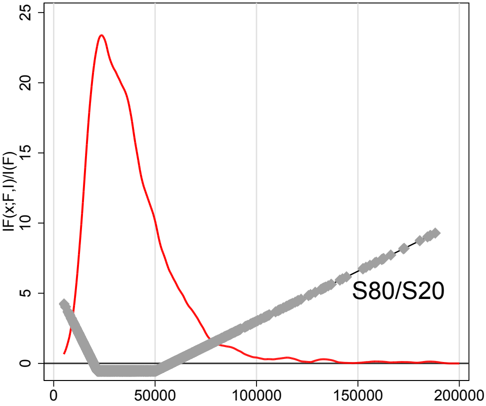


(annual household income data for Luxembourg 2012)

Income inequality: other measures

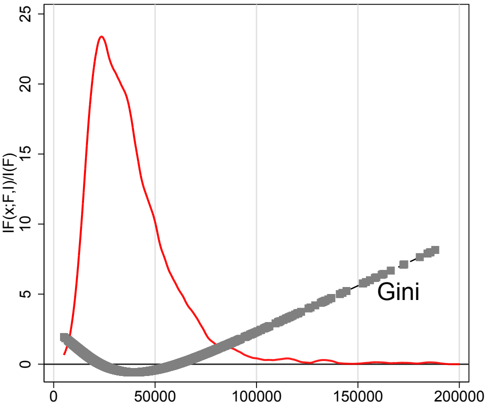
How does it compare with the other possible indicators of inequality:

Quintile (Group) Share Ratio?



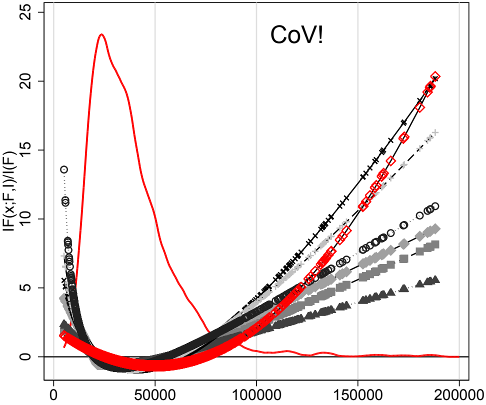
Income inequality: other measures

How does it compare with the other possible indicators of inequality: Gini coefficient?



Income inequality: other measures

How does it compare with the other possible indicators of inequality ...



Practical use 2

Practical use 2:

- estimation of the sampling variance of the index
- asymptotic approximation that works with complex non-linear statistics
- works seamlessly 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= ...  // point estimate is added to IF eval
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
```

Variance estimation

svyset as usual

```
. svyset uorigid [pw=wvar]  , strata(ustrata)  singleunit(centered)

      pweight: wvar
          VCE: linearized
  Single unit: centered
     Strata 1: ustrata
         SU 1: uorigid
        FPC 1: <zero>
```

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   =         63                  Number of obs     =      3,731
Number of PSUs     =      3,679                  Population size   = 511,900.22
                                                 Design df         =      3,616
                                                 F(    0,   3616)  =          .
                                                 Prob > F          =          .

------------------------------------------------------------------------------
             |             Linearized
       nivie |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
atkp5        |
       _cons |   .0634366   .0022941    27.65   0.000     .0589388    .0679345
-------------+----------------------------------------------------------------
atk1         |
       _cons |   .1211497   .0039797    30.44   0.000     .1133469    .1289524
-------------+----------------------------------------------------------------
atk2         |
       _cons |   .2228505   .0064922    34.33   0.000     .2101217    .2355793
-------------+----------------------------------------------------------------
s80s20       |
       _cons |   4.092545   .0973684    42.03   0.000     3.901642    4.283447
------------------------------------------------------------------------------
```

Variance estimation

Built in some user-written commands

```
. svy , subpop(if chme11==1) : inequaly nivie  , atkinson(0.5 1 2)  bonferroni
(running inequaly on estimation sample)

Survey data analysis

Number of strata   =         63                  Number of obs     =      3,731
Number of PSUs     =      3,679                  Population size   = 511,900.22
                                                 Subpop. no. obs   =      2,337
                                                 Subpop. size      = 282,251.64
                                                 Design df         =      3,616
                                                 F(   0,   3616)   =          .
                                                 Prob > F          =          .

------------------------------------------------------------------------------
             |             Linearized
       nivie |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
atkp5        |
       _cons |   .0544238   .0028886    18.84   0.000     .0487603    .0600872
-------------+----------------------------------------------------------------
atk1         |
       _cons |   .1050319   .0050275    20.89   0.000     .0951749     .114889
-------------+----------------------------------------------------------------
atk2         |
       _cons |   .1991793   .0084773    23.50   0.000     .1825586       .2158
-------------+----------------------------------------------------------------
bon          |
       _cons |   .3600872   .0070422    51.13   0.000     .3462802    .3738942
------------------------------------------------------------------------------
```

Practical use 3

Practical use 3:

- ‘Recentered IF regression’ (Firpo et al., 2007, 2009)
  - evaluate impact of covariates on statistics of interest
  - or what covariates are associated with large ‘influence’?
  - ‘unconditional’ (as in Firpo et al.’s ‘Unconditional quantile regressions’)

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 $v(F)$ 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 $v(F)$ 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

Effect of foreign households on inequality and poverty?

- Panel Study Liewen zu Letzebuerg 2011 (official source for poverty and inequality statistics in Luxembourg)
- Effect of a marginal increase in share of foreign-headed households on 'social indicators'
  - Assuming no change in income structure otherwise...
  - ... but conditioning on age of foreign households

Illustrative example

Code

```
svy:  inequaly nivie , atkinson(0.5 1 2)
predict rif* , rif // predict after -inequaly- gives (R)IF
svy:  regress rif1 i.(chme11)
svy:  regress rif2 i.(chme11)
svy:  regress rif3 i.(chme11)
svy:  inequaly nivie , s80s20
predict rifs80s20 , rif
svy:  regress rifs80s20 i.(chme11) i.(chme09)
svy:  newpoverty nivie , fracmedian(.6)
predict rifh , rif
svy:  regress rifh ib9.(rot)
```

Results: Atkinson(0.5)

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

Survey: Linear regression

Number of strata   =        63            Number of obs     =      3,704
Number of PSUs     =     3,653            Population size   = 506,643.98
                                          Design df         =      3,590
                                          F(   3,    3588)  =       6.94
                                          Prob > F          =     0.0001
                                          R-squared         =     0.0088

------------------------------------------------------------------------------
             |             Linearized
        rif1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      chme11 |
   Portugais |   .0120889    .0046899     2.58   0.010     .0028938     .021284
Autres UE-15 |   .0091529     .006223     1.47   0.141    -.0030481    .0213538
  Non UE-15  |   .0389485     .008964     4.34   0.000     .0213734    .0565236
             |
       _cons |   .0568863    .0034404    16.53   0.000     .0501409    .0636316
------------------------------------------------------------------------------
```

Results: Atkinson(0.5)

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

Survey: Linear regression

Number of strata   =         63                  Number of obs     =      3,704
Number of PSUs     =      3,653                  Population size   = 506,643.98
                                                 Design df         =      3,590
                                                 F(   8,    3583)  =     128.91
                                                 Prob > F          =     0.0000
                                                 R-squared         =     0.0124

------------------------------------------------------------------------------
             |             Linearized
        rif1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      chme11 |
   Portugais |   .0109198    .0048638     2.25   0.025     .0013837    .0204558
Autres UE-15 |    .009008    .0062075     1.45   0.147    -.0031626    .0211786
   Non UE-15 |   .0381033    .0090639     4.20   0.000     .0203324    .0558742
             |
      chme09 |
         <16 |  -.0440886     .004096   -10.76   0.000    -.0521194   -.0360578
     [16-24] |   .0366149    .0162167     2.26   0.024       .00482    .0684098
     [25-34] |   .0058699    .0057638     1.02   0.309    -.0054307    .0171705
     [35-49] |   .0123269    .0055662     2.21   0.027     .0014136    .0232402
     [50-64] |   .0171677    .0068698     2.50   0.012     .0036986    .0306368
             |
       _cons |   .0459795     .004096    11.23   0.000     .0379487    .0540103
------------------------------------------------------------------------------
```

Results: QSR

```
.         svy: regress rifs80s20 i.(chme11) ib6.(chme09)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         63                  Number of obs     =      3,704
Number of PSUs     =      3,653                  Population size   = 506,643.98
                                                 Design df         =      3,590
                                                 F(   8,    3583)  =      67.63
                                                 Prob > F          =     0.0000
                                                 R-squared         =     0.0113

------------------------------------------------------------------------------
             |             Linearized
   rifs80s20 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      chme11 |
   Portugais |   -.188026   .2295635    -0.82   0.413    -.6381139     .262062
Autres UE-15 |   .3467313   .2627993     1.32   0.187    -.1685195    .8619821
  Non UE-15  |   1.429164   .5102629     2.80   0.005     .4287298    2.429598
             |
      chme09 |
        <16  |  -1.580931   .1790653    -8.83   0.000    -1.932011   -1.229851
     [16-24] |   1.471687   .8037972     1.83   0.067    -.1042585    3.047631
     [25-34] |   .5860454   .3011544     1.95   0.052    -.0044055    1.176496
     [35-49] |   .5045821   .2424799     2.08   0.038     .0291698    .9799943
     [50-64] |   .6556552   .2622047     2.50   0.012      .1415701     1.16974
             |
       _cons |   3.477878   .1790653    19.42   0.000     3.126798    3.828958
------------------------------------------------------------------------------
```

Results: Poverty rate

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

Survey: Linear regression

Number of strata   =         63                 Number of obs      =      3,704
Number of PSUs     =      3,653                 Population size    = 506,643.98
                                                Design df          =      3,590
                                                F(   8,    3583)   =      61.12
                                                Prob > F           =     0.0000
                                                R-squared          =     0.0128

------------------------------------------------------------------------------
             |             Linearized
        rifh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      chme11 |
   Portugais |  -.0867281    .0291859    -2.97   0.003    -.1439508   -.0295054
Autres UE-15 |  -.0101697    .0218373    -0.47   0.641    -.0529845     .032645
   Non UE-15 |   .0167683    .0478889     0.35   0.726     -.077124    .1106605
             |
      chme09 |
         <16 |   .1970039    .0150227    13.11   0.000       .16755    .2264578
     [16-24] |   .1299979     .113194     1.15   0.251     -.091933    .3519288
     [25-34] |   .0949819    .0300992     3.16   0.002     .0359687     .153995
     [35-49] |   .0478833    .0214961     2.23   0.026     .0057374    .0900292
     [50-64] |    .048386    .0207141     2.34   0.020     .0077734    .0889987
             |
       _cons |   .1260334    .0150227     8.39   0.000     .0965795    .1554874
------------------------------------------------------------------------------
```

Results: Poverty rate (fixed poverty line)

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

Survey: Linear regression

Number of strata   =         63                  Number of obs     =      3,704
Number of PSUs     =      3,653                  Population size   = 506,643.98
                                                 Design df         =      3,590
                                                 F(   8,    3583)  =      45.08
                                                 Prob > F          =     0.0000
                                                 R-squared         =     0.0998

------------------------------------------------------------------------------
             |             Linearized
    rifhbis1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      chme11 |
   Portugais |   .2125625   .0291885     7.28   0.000     .1553348    .2697903
Autres UE-15 |   .0120222   .0195524     0.61   0.539    -.0263126    .0503571
   Non UE-15 |   .2905906    .050822     5.72   0.000     .1909477    .3902336
             |
      chme09 |
         <16 |  -.0369874   .0092244    -4.01   0.000     -.055073   -.0189019
     [16-24] |   .3534741   .1375077     2.57   0.010     .0838729    .6230752
     [25-34] |   .0904637   .0291149     3.11   0.002     .0333803    .1475471
     [35-49] |   .0644081   .0165652     3.89   0.000       .03193    .0968862
     [50-64] |   .0499869   .0156896     3.19   0.001     .0192254    .0807484
             |
       _cons |   .0369874   .0092244     4.01   0.000     .0189019     .055073
------------------------------------------------------------------------------
```

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.

Cowell, F. A. and Flachaire, E. (2007), 'Income distribution and inequality measurement: The problem of extreme values', *Journal of Econometrics* **141**(2), 1044–1072.

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’).