

# Estimation error due to duplicated observations: a Monte Carlo simulation.

Francesco Sarracino & Małgorzata Mikucka

*Statistical Office of Luxembourg, University of Leuven-la-Neuve, &  
Higher School of Economics*

Wednesday 20<sup>th</sup> April, 2016

## A question for you



# Statistical consequences of non-unique observations

## Literature review

# Statistical consequences of non-unique observations

## Literature review



## What does it mean?



# Definition

## Duplicate records are:

records that are not unique, i.e. records in which the set of all (or nearly all) answers from a given respondent is identical to that on another respondent.

They originate from:

- ▶ error or forgery by interviewers;
- ▶ data coders;
- ▶ data processing staff.

*(American Statistical Association, 2003; Kuriakose & Robbins, 2015; Waller, 2013)*

# How frequent are duplicated data?

We assume that the data we use are reliable, but . . .

“non-unique records occur at non-negligible rates” (*Kuriakose & Robbins, 2015*).

- ▶ Slomczynski et al. 2015: considerable amount of duplicates in 17/22 international surveys;
- ▶ Kuriakose & Robbins, 2015: 20% of 1000 public datasets contain duplicated observations.

# It seems an important topic

## Slomczynski et al., 2015

1721 national surveys from 22 comparative survey projects, 142 countries, 2.3 millions respondents:

- ▶ Surveys with duplicates **are frequent**: ISSP (35.8%); LatinoBarometro (68.4%); WVS (19.6%); ESS (3.4%).
- ▶ Duplicates **are not many**: on average no more than 1% duplicate records (sometimes  $> 10\%$ ).
- ▶ Duplicates come with **various patterns**:
  - ▶ Ecuador (2000) in Latinobarometro: 60% of duplicate records (doublets (272), triplets (63));
  - ▶ Norway (2009) in ISSP: 11% of duplicate records (doublets (27), triplets (12), quadruplets (6), quintuplets (5), and more.)



# It seems an important topic

## Kuriakose & Robbins, 2015

1008 national surveys, more than 1.2 million observations, 35 years, 154 countries, territories or subregions:

- ▶ 20% of the surveys has duplicated data;
- ▶ 30% of 309 of Pew's international studies has duplicated data;
- ▶ in Western countries 5% of the surveys have duplicated data;
- ▶ in the developing world, it's 26%.
- ▶ only rarely non-unique cases are identical on all variables (*near duplicates*).

How the debate is going:



# “The problem isn’t going to just go away”

The question remains: “how duplicate records affect results of regression analysis, and to deal with them?”

# “The problem isn't going to just go away”

The question remains: “how duplicate records affect results of regression analysis, and to deal with them?”



## Our contribution

We assess **the risk of obtaining biased estimates due to duplicated observations**:

Duplicate cases:

- ▶ increase the sample used in statistical inference;
- ▶ reduce the variance;
- ▶ artificially increase statistical power of estimations;
- ▶ narrower estimated confidence intervals

Risk of getting wrong conclusions!

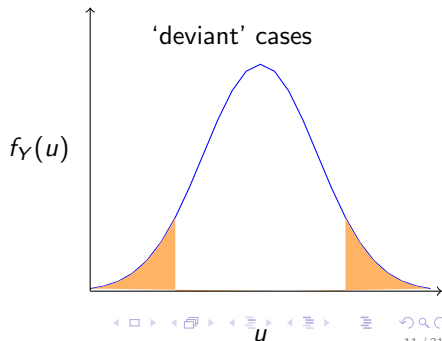
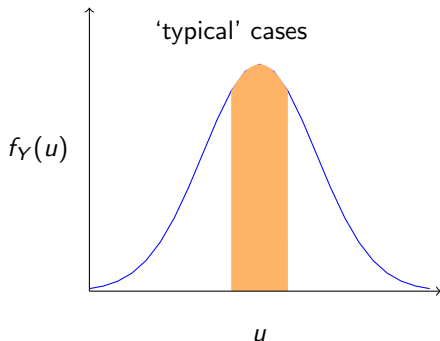
## Our contribution

We assess **the risk of obtaining biased estimates due to duplicated observations:**

Duplicate cases:

- ▶ increase the sample used in statistical inference;
- ▶ reduce the variance;
- ▶ artificially increase statistical power of estimations;
- ▶ narrower estimated confidence intervals

Risk of getting wrong conclusions!



# Our contribution

We assess the **reliability of possible solutions**:

- ▶ naive estimation;
- ▶ dropping the duplicate observations;
- ▶ flagging the duplicate observations;
- ▶ robust regression;
- ▶ weighting for the inverse of the multiplicities.

## How we do it





## How we do it



### Monte Carlo simulation:

how various **numbers** and **patterns** of duplicate records affect the risk of obtaining biased estimates.

## We proceed in 4 steps:

### 1) we generate the initial dataset:

- ▶  $N = 1500$
- ▶ Variables:  $x$ ,  $y$ ,  $z$ , and  $t$ ;  $y$  is treated as dependent variable;
- ▶ Matrix of correlations used to generate the original dataset.

variables	x	y	z	t
x	1			
y	0.50	1		
z	0.40	0.94	1	
t	-0.43	-0.81	-0.80	1

- ▶ *true* coefficients:  $y_i = \alpha + \beta_1 \cdot x_i + \beta_2 \cdot t_i + \beta_3 \cdot z_i + \varepsilon_i$

# We proceed in 4 steps:

## 2) we duplicate randomly selected cases:

- ▶ Monte Carlo simulation to generate duplicate records and to replace original ones;

### Scenario 1

a single observation is duplicated from 1 to 5 times:

#### Variant 1

duplicate records are chosen randomly

#### Solution 1

'naive' estimation

#### Variant 2

duplicate records are from the second and third quartile

#### Solution 2

excluding all duplicate records

#### Solution 3

flagging the duplicate records

### Scenario 2

data contain multiple pairs of identical records (1 - 79 doublets):

#### Variant 3

duplicate records are from the lower quartile

#### Solution 4

robust regression VS OLS

#### Variant 4

duplicate records are from the upper quartile

#### Solution 5

weighting by the inverse of the multiplicity

- ▶ We investigate 40 patterns ( $2 \cdot 4 \cdot 5 = 40$ ) of duplicate records.
- ▶ For each pattern we run 1000 repetitions in which duplicated and replaced records are chosen randomly according to the variants.

## We proceed in 4 steps:

### 3) 'naive' estimation and possible solutions:

- ▶ 'naive' estimation: takes data as they are;
- ▶ excluding duplicate records;
- ▶ flagging duplicate records and control for them;
- ▶ robust regression: duplicate records constitute influential observations and we can account for this;
- ▶ weighting by the inverse of multiplicities (*Lessler & Kalsbeek, 1992*).

## We proceed in 4 steps:

### 4) assessment of bias:

- ▶ we subtract the 'true' coefficients from those estimated for data with duplicates;
- ▶ we use  $Dfbetas$  to assess the severity of the bias;

### What are $Dfbetas$ ?

Normalized measures of how much specific observations affect the estimates of regression coefficients.

$$Dfbeta = \frac{\beta_{new} - \beta_{true}}{se_{new}}$$

High bias if  $Dfbetas > \frac{2}{\sqrt{N}} = 0.05$ .

## An example of the dataset produced in a repetition

N. of duplicates	variable	mean	sd	min	max	obs	missing
Initial dataset		3016	749.7	344.9	5775	1500	0
		6176	2899	-3213	17299	1500	0
		187.8	21.71	103.2	261.4	1500	0
		21.25	5.633	1.967	41.45	1500	0
	duplicates (flag)	0	0	0	0	1500	0
1 doublet		3015	750.0	344.9	5775	1500	0
		6176	2899	-3213	17299	1500	0
		187.8	21.71	103.2	261.4	1500	0
		21.25	5.633	1.967	41.45	1500	0
	duplicates (flag)	0.000667	0.0258	0	1	1500	0
1 triplet		3017	748.9	344.9	5775	1500	0
		6177	2898	-3213	17299	1500	0
		187.8	21.68	103.2	261.4	1500	0
		21.24	5.627	1.967	41.45	1500	0
	duplicates (flag)	0.00133	0.0365	0	1	1500	0
1 quadruplet		3018	753.5	344.9	5775	1500	0
		6183	2902	-3213	17299	1500	0
		187.9	21.80	103.2	261.4	1500	0
		21.23	5.657	1.967	41.45	1500	0
	duplicates (flag)	0.00200	0.0447	0	1	1500	0
1 quintuplet		3017	748.3	344.9	5775	1500	0
		6180	2895	-3213	17299	1500	0
		187.8	21.66	103.2	261.4	1500	0
		21.24	5.630	1.967	41.45	1500	0
	duplicates (flag)	0.00267	0.0516	0	1	1500	0
1 sextuplet		3014	747.6	344.9	5775	1500	0
		6175	2893	-3213	17299	1500	0
		187.7	21.67	103.2	261.4	1500	0
		21.27	5.624	1.967	41.45	1500	0
	duplicates (flag)	0.00333	0.0577	0	1	1500	0

Are you still with me?



# To recap

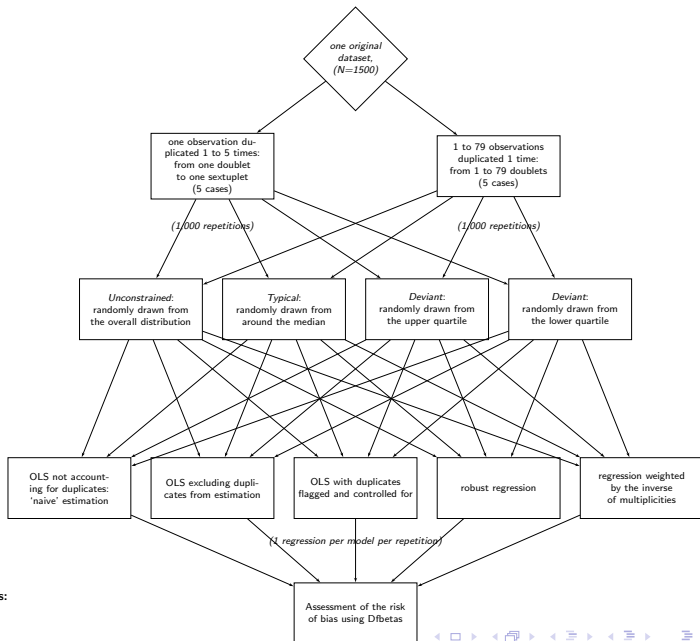
Data generation:

Scenarios:

Variants:

Solutions:

Assessment of bias:

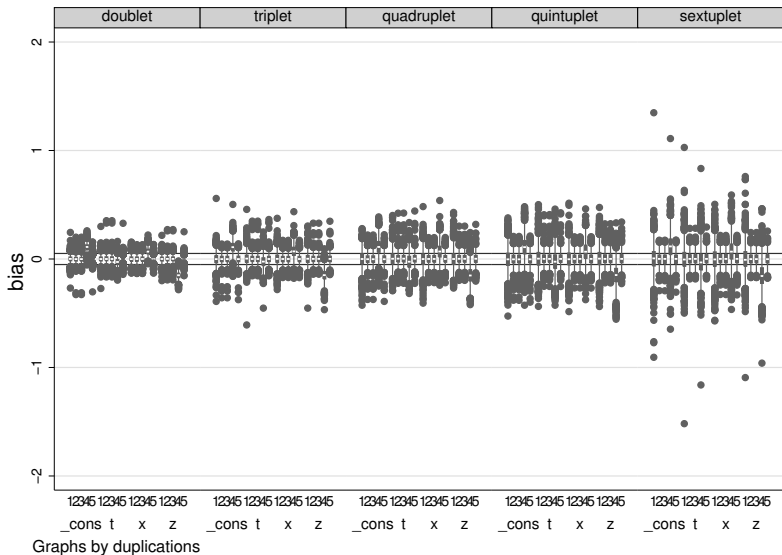




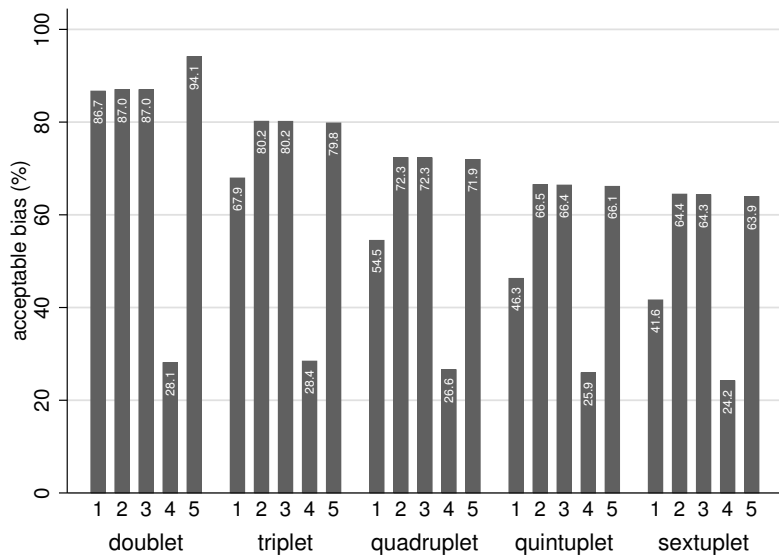
## What we have found



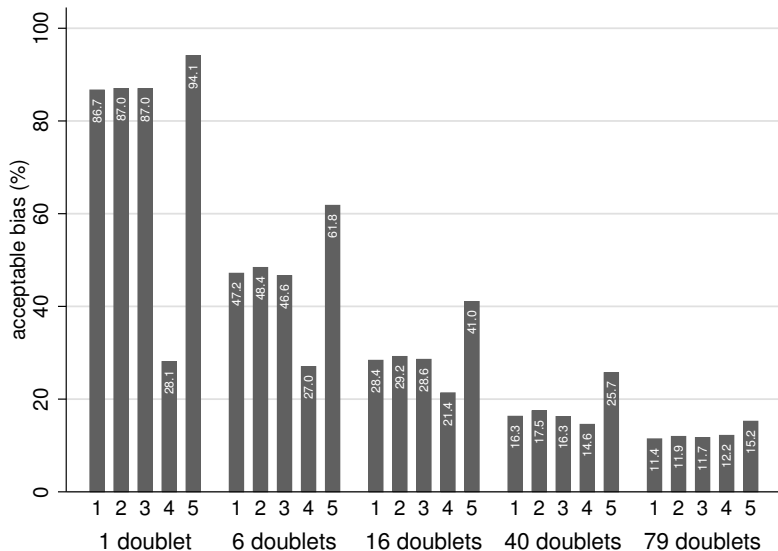
# Errors when 1 observation is duplicated 1 to 5 times.



# Probability of obtaining unbiased coefficients.



# Probability of unbiased coefficients when 1 to 79 obs. are duplicated 1 time.



## First conclusions

- ▶ Weighting for the inverse of the multiplicities **decreases the risk** of obtained erroneous estimates if 1 doublet is present;
- ▶ Dropping, flagging and weighting work well when data have a single triplet, quadruplet, quintuplet or sextuplet;
- ▶ Dropping and flagging **perform poorly** if multiple doublets are included in the data;
- ▶ Robust regression **performs poorly** in all cases.

## Typical and deviant cases



Are the risks of obtaining wrong estimates lower if the duplicate records are 'typical'?

# Typical and deviant cases: 1 obs. duplicated many times

	Duplicated observation drawn randomly from:			
	overall distribution	center of distribution	lower quartile	upper quartile
<i>1 doublet:</i>				
'Naive' estimation	86.67	88.22	87.17	85.40
Drop duplicates	87	86.13	87.10	86.53
Flag and control	86.97	86.13	87.10	86.53
Robust regression	28.10	27.80	26.77	26.40
Weighted regression	94.10	93.63	94.03	94.05
<i>1 quadruplet:</i>				
'Naive' estimation	54.48	53.38	55.30	55.17
Drop duplicates	72.33	71.30	74.20	75.22
Flag and control	72.28	71.22	74.20	75.15
Robust regression	26.57	25.55	29.65	25.73
Weighted regression	71.92	70.90	73.90	74.72
<i>1 sextuplet:</i>				
'Naive' estimation	41.63	39.60	39.23	39.40
Drop duplicates	64.45	64.72	63.13	61.90
Flag and control	64.30	64.60	62.95	61.85
Robust regression	24.18	22.50	24.70	23.75
Weighted regression	63.90	64.30	62.63	61.58

## Second conclusion

- ▶ 'typical' or 'deviant' cases make little difference for the risk of getting wrong estimates;
- ▶ the risk of error when the duplicate is drawn from the overall distribution **is not lower** than when the duplicate is drawn from the tie.
- ▶ these results do not depend on the solution adopted to deal with duplicates.
- ▶ These results generally hold also when many observations are duplicated once.
- ▶ These conclusions do not change if the duplicate records are drawn on the basis of the distribution of the  $x$  variable.



# Concluding remarks

## Be aware that duplicate records affect your estimates!!!

- ▶ The risk of obtaining wrong estimates increases with the number of duplicate records:
  - ▶ a single sextuplet ( $< 1\%$ ) the probability of unbiased estimates is 41.6%;
  - ▶ 79 doublets of identical records ( $\sim 10\%$ ) the probability of unbiased estimates is 11.4%.
- ▶ Even a small number of duplicate records creates considerable risk of wrong estimates.
- ▶ The risk of wrong estimates does not change for 'typical' and 'deviant' cases;
- ▶ Weighting the duplicates by the inverse of their multiplicity is the best solution (among the considered ones) to minimize the risk of wrong estimates.

## Policy recommendation



## Policy recommendation



- ▶ It is possible to adopt solutions to minimize the errors;
- ▶ Correcting the data with statistical tools is not a trivial task.

**Thanks a lot for your attention!**

**Francesco.Sarracino@statec.etat.lu**  
**f.sarracino@gmail.com**

This report was presented at the 6th LCSR International Workshop  
“Trust, Social Capital and Values in a Comparative Perspective”,  
which held within the XVII April International Academic Conference on Economic and Social Development.

April 18 – April 22, 2016 - Higher School of Economics, Moscow.

<https://lcsr.hse.ru/en/seminar2016>

Настоящий доклад был представлен на VI международном рабочем семинаре ЛССИ  
«Доверие, социальный капитал и ценности в сравнительной перспективе»,  
прошедшего в рамках XVII Апрельской международной научной конференции НИУ ВШЭ «Модернизация экономики и общества».

18 – 22 апреля, 2016 – НИУ ВШЭ, Москва.

<https://lcsr.hse.ru/seminar2016>