Estimation bias due to duplicated observations: a Monte Carlo simulation

Sarracino, Francesco and Mikucka, Malgorzata

Institut national de la statistique et des études économiques du Grand-Duché du Luxembourg (STATEC), Université catholique de Louvain, National Research University Higher School of Economics, Laboratory for Comparative Social Research

26 January 2016

Estimation bias due to duplicated observations: a Monte Carlo simulation*

Francesco Sarracino† and Małgorzata Mikucka‡

January 26, 2016

Abstract

This paper assesses how duplicate records affect the results from regression analysis of survey data, and it compares the effectiveness of five solutions to minimize the risk of obtaining biased estimates. Results show that duplicate records create considerable risk of obtaining biased estimates. The chances of obtaining unbiased estimates in presence of a single sextuplet of identical observations is 41.6%. If the dataset contains about 10% of duplicated observations, then the probability of obtaining unbiased estimates reduces to nearly 11%. Weighting the duplicate cases by the inversion of their multiplicity minimizes the bias when multiple doublets are present in the data. Our results demonstrate the risks of using data in presence of non-unique observations and call for further research on strategies to analyze affected data.

Key-words: duplicated observations, estimation bias, Monte Carlo simulation, inference.

1 Introduction

The work of applied researchers often relies on survey data, and the reliability of the results depends on accurate recording of respondents' answers. Yet, sometimes this condition is not met. A recent study by Slomczynski et al. (2015) investigated survey projects which are widely used in social sciences, and reported a considerable number of duplicate records in 17 out of 22 projects. Duplicate records are defined as records that are not unique, that is records in which the set of all (or nearly all) answers from a given respondent is identical to that of another respondent.

Surveys in social sciences usually include a large number of questions. Thus, it is highly unlikely that two respondents provided identical answers to all (or nearly all) substantive survey questions. In other words, it is unlikely that two identical records come from answers of two real respondents. It is more probable that either one record corresponds to a real respondent and the second one is its duplicate, or that both records are fake. Duplicate records can result from an error or forgery by interviewers, data coders, or data processing staff and should, therefore, be treated as suspicious observations (American Statistical Association, 2003; Kuriakose and Robbins, 2015; Waller, 2013).

This is the first analysis assessing how duplicate records affect estimated regression coefficients. We investigate two scenarios of duplicate data: when one record is duplicated several times (bias due to the number of duplications), and when more than one record is duplicated two times (bias due to number

*The authors wish to thank Kazimierz M. Slomczynski, Przemek Powałko, and the participants to the Harmonization Project of the Polish Academy of Science for their comments and suggestions. Possible errors or omissions are entirely the responsibility of the authors who contributed equally to this work.

†Institut national de la statistique et des études économiques du Grand-Duché du Luxembourg (STATEC), and LCSR National Research University Higher School of Economics, Russian Federation. 13, rue Erasme, L-2013, Luxembourg. Tel.: +352 247 88469; e-mail: f.sarracino@gmail.com

‡Université catholique de Louvain and LCSR National Research University Higher School of Economics, Russian Federation.

of duplicated records). We also consider five solutions to deal with duplicate cases, and we assess their ability to reduce the estimation bias. We consider the following solutions: excluding the duplicate cases from the analysis, flagging the duplicate cases and including the flags in the model, using robust regression model as a way to minimize the effect of influential observations, and weighting the duplicate cases by the inverse of their multiplicity.

1.1 Non-unique cases in social survey data

Slomczynski et al. (2015) analyzed a set of 1,721 national surveys belonging to 22 comparative survey projects, with data coming from 142 countries and nearly 2.3 million respondents. The analysis identified 5,893 non-unique records in 162 national surveys from 17 projects coming from 80 countries. The duplicate records were unequally distributed across the surveys. For example, they appeared in 19.6% of surveys of the World Values Survey (waves 1-5) and in 3.4% of surveys of the European Social Survey (waves 1-6). Across survey projects, also different numbers of countries were affected. Latinobarometro is an extreme case where surveys from 13 out of 19 countries contained non-unique records. In the Americas Barometer 10 out of 24 countries were affected, and in the International Social Survey Programme 19 out of 53 countries contained duplicate cases.

Even though the share of duplicate records in most surveys did not exceed 1%, in some of the national surveys it was high, exceeding 10% of the sample. In 52% of the affected surveys Slomczynski et al. (2015) found only a single pair of non-unique records. However, in 48% of surveys containing duplicates they found various patterns of non-unique records, such as multiple doublets (i.e. multiple pairs of identical records) or identical records repeated three, four, or more times. For instance, Slomczynski et al. (2015) identified 733 non-unique records (60% of the sample), including 272 doublets and 63 triplets in the Ecuadorian sample of Latinobarometro collected in the year 2000. Another example are data from Norway registered by the International Social Survey Programme in 2009, where 54 non-unique records consisted of 27 doublets, 36 non-unique records consisted of 12 triplets, 24 consisted of 6 quadruplets, 25 consisted of 5 quintuplets; along with, one sextuplet, one septuplet, and one octuplet (overall 160 non-unique records, i.e. 11.0% of the sample).

Note that only rarely the non-unique cases are identical on all variables. In most cases repeated records are *near duplicates*, i.e. duplicate observations in which responses differ for only a small number of variables. This might be the result of forgery in which some responses have been changed to avoid detection by software running tests for exact duplicates. Kuriakose and Robbins (2015) analyzed near duplicates in datasets commonly used in social sciences. They stressed that demographic and geographic variables are rarely falsified, because they typically need to meet the sampling frame. Behavioral and attitudinal variables are likely falsified more often. In such cases copying only selected sequences of answers provided by some respondents may ensure that the achieved correlations between variables are as expected, and the forgery remains undetected. In the analysis by Kuriakose and Robbins (2015), 16% of analyzed surveys revealed a high risk of widespread falsification with near duplicates.

1.2 Current analysis

Confronted with these results, practitioners may want to know how duplicate records affect results of regression analysis, and how to deal with them. Duplicate cases increase the sample used in statistical inference, reduce the variance, and thus they may artificially increase statistical power of estimation methods. This may result in narrower estimated confidence intervals, thus solidifying the estimated relationships between variables (Kuriakose and Robbins, 2015). This may lead to more significant coefficients, thus affecting the conclusions from the studies. The risk may be particularly high when duplicate records are 'deviant' cases, as they may influence estimation procedures more than 'typical' cases. Slomczynski et al. (2015) refer to 'typical' cases as duplicate records located near the median of the relevant variable, while they refer to 'deviant' cases as duplicate records located close to the ties of the distribution.

However, these are merely speculations because the literature on this topic is virtually not existing, and focuses mainly on strategies to identify duplicate and near-duplicate records (Elmagarmid et al., 2007; Kuriakose and Robbins, 2015).

The goal of our analysis is to assess the risk of obtaining biased estimates due to duplicated observations. We use a Monte Carlo simulation to investigate how various numbers and patterns of duplicate records affect the risk of obtaining biased estimates. We consider two main scenarios: the first when a single observation is duplicated between one and five times (thus including in the data a single doublet, triplet, quadruplet, quintuplet, and sextuplet of identical observations), and the scenario when data contain multiple pairs of identical records (between 1 and 79 doublets).

We also investigate how the risk of bias changes when duplicates are located in specific parts of the distribution of the dependent as well as of one of the independent variables. We expect that duplicate records bias estimates more if they are 'deviant' i.e. when they take values far away from the median. To this end, we evaluate the bias in four variants, namely:

- when the duplicate records are chosen randomly from the whole distribution of the dependent variable (we label this variant 'unconstrained' as we do not impose any limitation on where the duplicate records are located);
- when they are chosen randomly from the second and third quartile of the dependent variable (i.e. when they are located around the median: this is the 'typical' variant);
- when they are chosen randomly from the lower quartile of the dependent variable (this is the first 'deviant' variant);
- when they are chosen randomly from the upper quartile of the dependent variable (this is the second 'deviant' variant).

We expect that duplicate cases drawn from the lower or upper quartiles bias the estimates more than duplicate records drawn from around the median, i.e. from the second and third quartile. To check the robustness of our findings, we follow the same scheme when we analyze how the position on the distribution of one of the independent variables affects the bias.

Finally, we evaluate five possible solutions to deal with the bias introduced by duplicate records. The solutions include: not accounting for duplicates – i.e. running a 'naive' estimation, excluding all duplicate records from the analysis, flagging the duplicates and controlling for them in the regression analysis, using robust regression instead of standard OLS estimation to diminish the effect of influential observations, and weighting the duplicate cases by the inverse of their multiplicity.

2 Method

To assess how duplicate records bias results of OLS regression we use a Monte Carlo simulation. Our analysis consists of four main steps. In the first step we generate the initial dataset. In the second step we duplicate randomly selected cases according to the two scenarios and four variants mentioned above. In the third step we estimate regression models using a 'naive' approach, i.e. treating data with duplicates as if they were correct; we also estimate regression models using five possible solutions to deal with duplicate cases. Finally, we compare the bias of estimates obtained from various scenarios of cases' duplication and we evaluate the effectiveness of the possible solutions. Figure 1 summarizes our strategy.

2.1 Data generation

We begin by generating a dataset with a known covariance matrix and vector of averages. The dataset contains $N = 1,500$ observations and four variables: x, y, z, and t. We treat variable y as the dependent variable, and variables x, z, and t as predictors. Table 1 shows the correlation matrix used to generate the original dataset.

Figure 1: Diagram summarizing the empirical strategy.

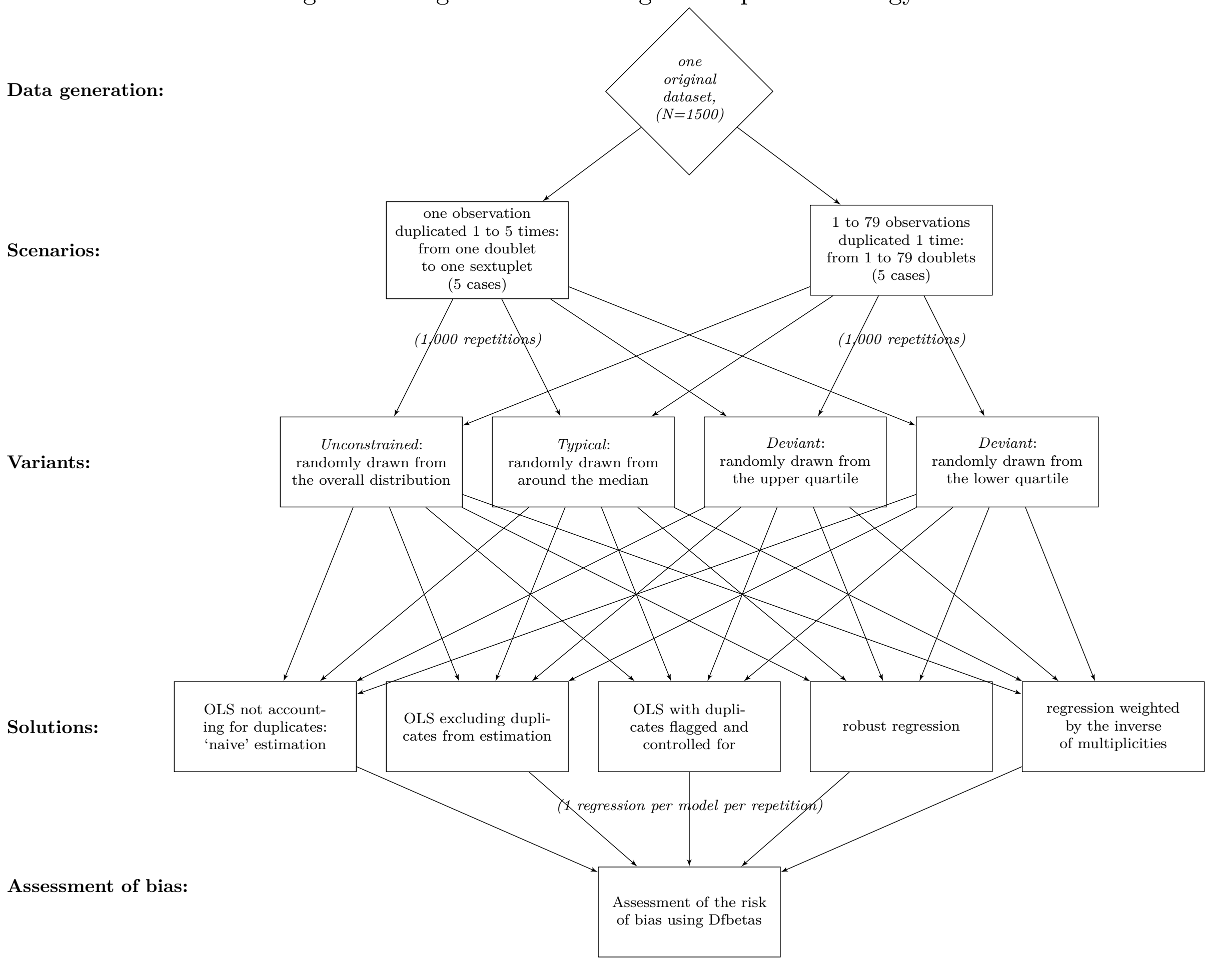


Table 1: Matrix of correlations used to generate the original dataset.

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

All variables are normally distributed. The descriptive statistics of the generated dataset are shown in Tables 4 and 5 in Appendix A.

After creating the dataset we estimate the 'true coefficients' of the relationships between the dependent variable y and the predictors in the absence of any duplicate case. We used a standard OLS regression as in Equation 1.

$$y_i = \alpha + \beta_1 \cdot x_i + \beta_2 \cdot t_i + \beta_3 \cdot z_i + \varepsilon_i \tag{1}$$

We treat the coefficients estimated from Equation 1 as the benchmarks to assess the bias of coefficients estimated for data with duplicate cases (see Section 2.4).

2.2 Duplicating selected cases

In the second step we use a Monte Carlo simulation to generate duplicate records, which replace some randomly chosen original records. This is motivated by the assumption that usually duplicate records substitute authentic interviews. Thus, if duplicates are present, practitioners do not only face the cost of fake or erroneous information, but they also lose information from genuine respondents.

We duplicate selected cases in two scenarios (each comprising five cases) and in four variants. Thus, overall we investigate 40 patterns ($2 \cdot 5 \cdot 4 = 40$) of duplicate records. For each pattern we run 1000 repetitions in which duplicated and replaced records are chosen randomly according to the variants detailed below, i.e. 'unconstrained', 'typical' and 'deviant'.

Scenario 1: from a doublet to a sextuplet In the first scenario we duplicate one randomly chosen record between 1 and 5 times, thus introducing in the data a doublet, a triplet, a quadruplet, a quintuplet, or a sextuplet of identical observations which replace for 1 to 5 original observations. These cases are plausible in the light of the analysis by Slomczynski et al. (2015), who identified in real survey data not only a sextuplet but even an octuplet. In all cases in this scenario the share of duplicates in the sample is small, ranging from 0.13% in case of a doublet to 0.40% in case of a sextuplet.

Scenario 2: from a single doublet to 79 doublets In the second scenario we duplicate sets of 1, 6, 16, 40, and 79 observations one time, creating between 1 and 79 pairs of identical observations replacing an equal number of original records. In this scenario the share of duplicate cases ranges from 0.13% (a single doublet, as in Scenario 1) to 10.5% (79 doublets). According to Slomczynski et al. (2015) this is a realistic number, as in their analysis about 15% of the affected surveys had 10% or more non-unique cases.

Variants: 'unconstrained', 'typical' and 'deviant' To check if the position of duplicates in the distribution matters, we run each of the scenarios in four variants:

1. **unconstrained**: randomly drawn from the overall distribution of the dependent variable;
2. **typical**: randomly drawn from the values around the median of the dependent variable, i.e. from the second and third quartile;
3. **deviant**: randomly drawn from the lower quartile of the dependent variable; and
4. **deviant**: randomly drawn from the upper quartile of the dependent variable.

As a robustness check we repeat our estimates duplicating cases according to the distribution of the x variable rather than the y. Results are consistent with those presented below.

2.3 'Naive' estimation and possible solutions

We begin with a 'naive' estimation which takes data as they are, and subsequently we investigate the remaining four solutions to deal with duplicates. For each of them we estimate the regression model as in Equation 1.

Solution 1: 'naive' estimation First, we investigate what happens when researchers neglect the presence of duplicate observations. In other words, we analyze data with duplicate records as if they were correct. This allows us to estimate the bias resulting from the mere presence of duplicate records as the difference from the 'true coefficients' (see Section 2.4).

Solution 2: Exclude duplicated observations Excluding the duplicate records from the dataset seems an obvious solution to the problem. "[E]liminating duplicate and near duplicate observations from analysis is imperative to ensuring valid inferences" (Kuriakose and Robbins, 2015, p. 2). As a rule, deleting duplicates would allow retaining a sample consisting of really interviewed respondents. However, this comes at the cost of missing information from genuine interviews, which were not conducted or were conducted partly, and which have been replaced by the duplicates.

Additionally, if records are identical on some, but not on all variables (most likely, differences may exist on demographic and geographical variables, which reflects the sampling scheme), it may be unclear which cases are the original ones, and which are fake duplicates which should be dropped. We also cannot exclude the possibility that all duplicate cases are forged and should be excluded from the data. Therefore, rather that deleting the forged cases and retaining the original ones, we investigate a solution that is readily available for all researchers, namely we exclude all non-unique records from the data.

Solution 3: Flag duplicated observations and control for them This solution is similar to the previous one in that we identify all non-unique records as suspicious. However, rather than excluding them, we generate a dichotomous variable (duplicate = 1, otherwise = 0), and include it in the estimations. By doing so we attempt to control for the error generated by the duplicate records, as suggested by Slomczynski et al. (2015).

Solution 4: Robust regression This solution assumes that duplicate records may constitute influential observations, i.e. observations whose deletion from the data may noticeably change the estimation results. For data containing influential observations, Stata statistical software offers the 'robust regression' tool as an alternative to the standard OLS regression (*rreg* command).

Robust regression in Stata is a form of weighted least squares regression. It assigns lower weights to influential observations. Thus, it may be seen as a compromise between entirely excluding duplicated observations from the analysis or keeping them in the data.

Solution 5: Weighting by the inverse of multiplicities This method has been proposed by Lessler and Kalsbeek (1992). We construct a weight which takes the values of 1 for unique records, and of the inverse of multiplicity for duplicate records. For example, our weight takes the value $\frac{1}{2}$ for doublets, $\frac{1}{3}$ for triplets, $\frac{1}{4}$ for quadruplets, etc. Subsequently, we estimate the OLS model with data weighted by these weights.

2.4 Assessment of bias

In the final step of the analysis we assess the estimation bias by subtracting the 'true coefficients' from those estimated for data with duplicates.

We resort to Dfbetas to assess the severity of the bias. Dfbetas are normalized measures of how much specific observations (in our case: duplicated observations) affect the estimates of regression coefficients. To calculate Dfbetas we first compute the difference between new and the 'true' coefficients and subsequently we divide the difference by the standard error of the new coefficient (see Equation 2).

$$Dfbeta = \frac{\beta_{new} - \beta_{true}}{se_{new}} \tag{2}$$

A bias is typically considered high if Dfbetas are larger than $\frac{2}{\sqrt{N}}$, where N is the sample size. With $N = 1,500$ the threshold value for Dfbetas is 0.05.

To illustrate our data, we report the descriptive statistics of some of the datasets produced during the repetitions in Appendix A. Table 4 reports figures related to the original dataset (first 5 rows) and to five datasets generated by including in the data between one doublet and one sextuplet. In a similar way, Table 5 shows the figures from the original dataset (first 5 rows) and from five datasets generated by including in the data between 1 and 79 doublets.

3 Results

3.1 Severity of the bias

Figure 2 shows the estimation bias for each coefficient for cases in Scenarios 1 (upper panel) and 2 (lower panel) for the 'unconstrained' variant, i.e. when the duplicate records are randomly drawn from the overall distribution of the dependent variable. It is apparent that the bias of coefficients increases with the number of duplications. In case of a single doublet Dfbetas reach the value of about 0.5, while for 79 doublets they approach the value of 1.

This result is made clearer by Figure 3 reporting the frequency with which the bias falls within the acceptable range, i.e. Dfbetas do not exceed the critical value of 0.05 (for all coefficients combined). The upper panel of Figure 3 confirms that the risk of the bias in a 'naive' estimation increases when we move from a doublet to a sextuplet. If a doublet is included in the data, 86.7% of 'naive' estimates are sufficiently close to the true coefficients. However, the presence of a triplet reduces the chances of unbiased estimates to 67.9%, a quadruplet reduces the chances to 54.5%, a quintuplet to 46.3%, and a sextuplet to 41.6%. In other words, even if only a single sextuplet of identical records is included in the data, i.e. much less than 1% of the sample, researchers have 58% chance of obtaining biased coefficients if duplicates are neglected.

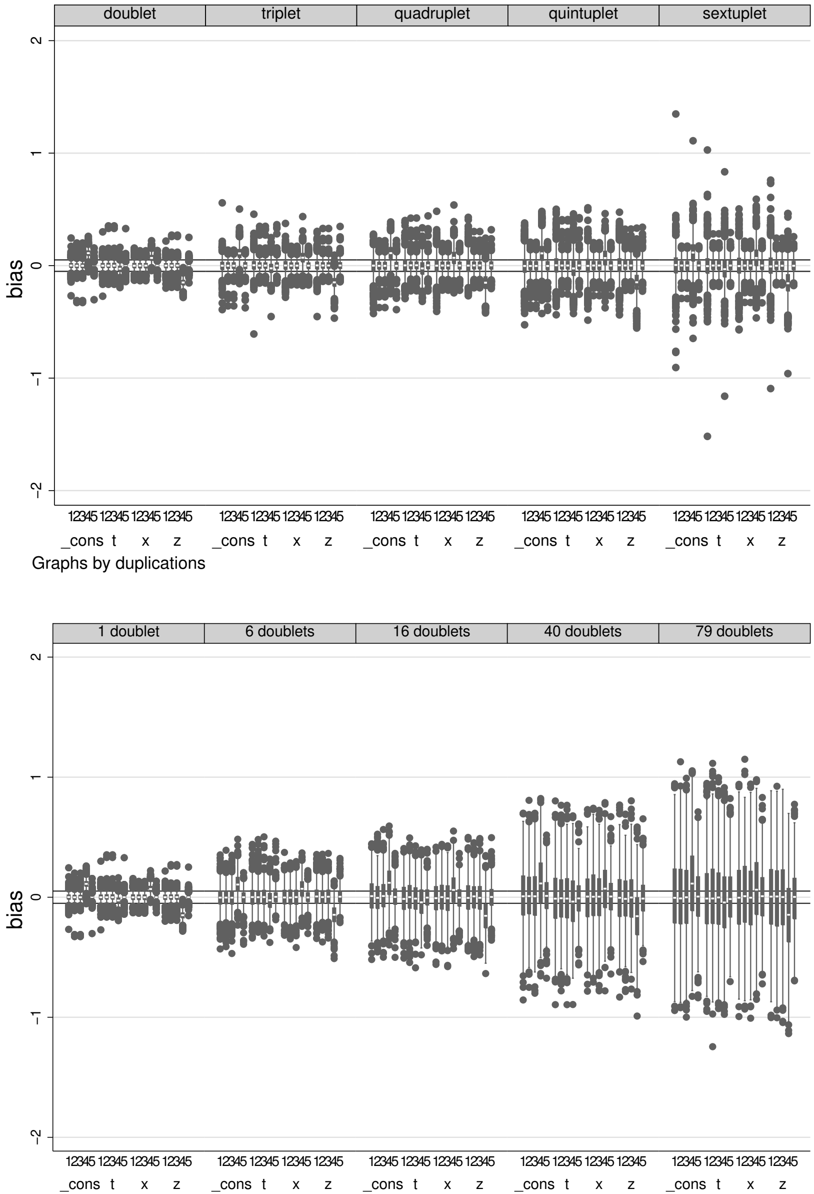
We observe a similar pattern when we increase the number of doublets included in the data (lower panel of Figure 3). In case of 'naive' estimation the probability of obtaining unbiased estimates decreases to 47.2% for 6 doublets, 28.4% for 16 doublets, 16.3% for 40 doublets, and 11.4% for 79 doublets. That is to say, when 79 pairs of duplicate records are present in the data, researchers ignoring the presence of duplicates have 89% chance of obtaining biased estimates.

3.2 Efficiency of solutions

After investigating the risk of bias in case of 'naive' estimation, we move to inspecting the performance of the proposed solutions. Upper panel of Figure 3 shows that, among the five considered solutions, only weighting for the inverse of the multiplicities decreases the risk of obtaining biased estimates if one doublet is present in the data. Additionally, excluding or flagging duplicated records (solutions coded as 2 and 3 in the graph) and weighting data by the inverse of multiplicity (solution 5) decrease the risk of obtaining biased estimates in data with a single triplet, quadruplet, quintuplet, or sextuplet. If a single triplet is excluded from the analysis, controlled for in the model, or weighted by the inverse of the multiplicity, the chance of obtaining unbiased estimates increases from 67.9% to about 79.8%, in case of a quadruplet − from 54.5 to about 71.2%, in case of a quintuplet − from 46.3 to 66.1%, and in case of a sextuplet − from 41.2 to 63.9%.

However, solutions 2 and 3 (i.e. excluding or flagging the duplicates) perform quite poorly if multiple doublets are included in the data (see the lower panel of Figure 3). The probability of obtaining biased estimates if duplicates are excluded or flagged is nearly the same as in the 'naive' estimation. Weighting by the inverse of the multiplicity performs relatively better in reducing the chances of biased estimates than other solutions. In this case the chances of unbiased estimates increase from 86.7% to 94.1% with

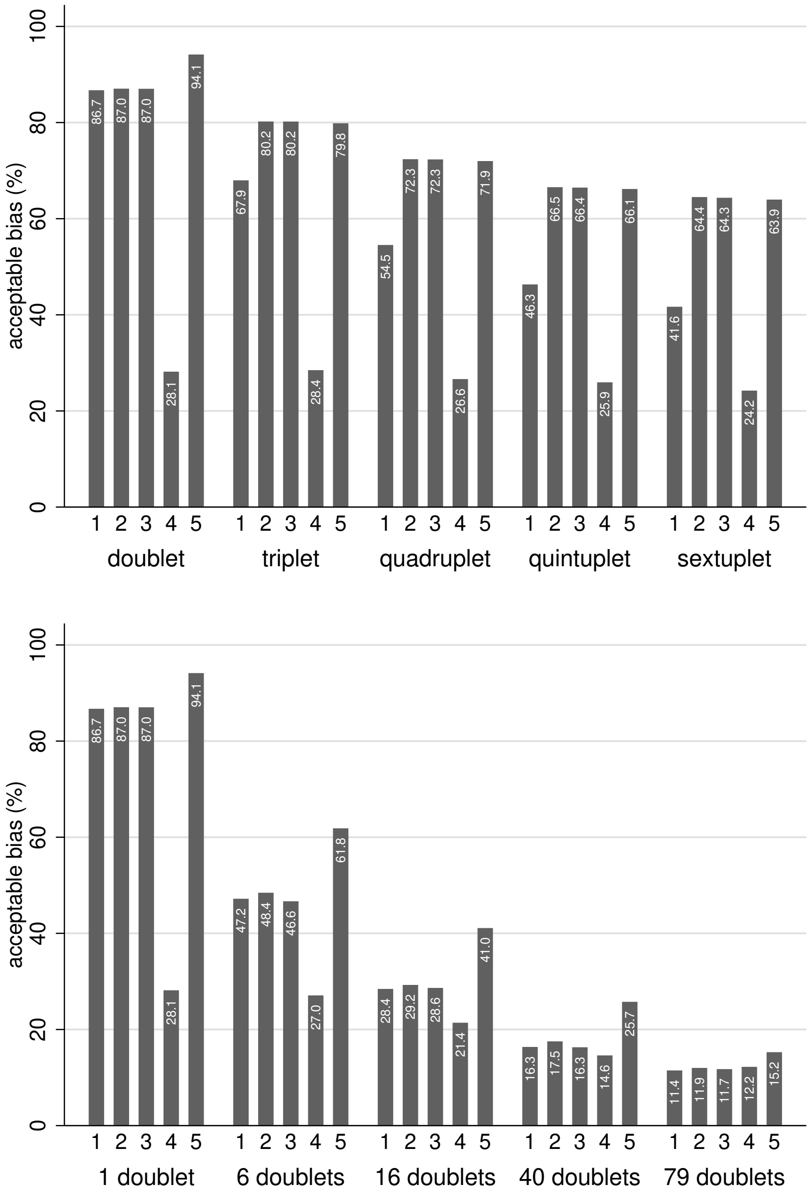
Figure 2: Bias (expressed in Dfbetas) for various scenarios of duplicate records.



1 – ‘naive’ estimation
2 – excluding the duplicate records
3 – duplicate records flagged and controlled for
4 – robust regressions
5 – weighted regressions

Notes: The duplicate records are randomly drawn from the overall distribution.
Box plots show the bias (expressed in Dfbetas) for each of the predictors in the model (constant, x, z and t) and for the five considered solutions (1 to 5).

Figure 3: Probability of obtaining unbiased coefficients for various scenarios of duplicate records.



1 – 'naive' estimation
2 – excluding the duplicate records
3 – duplicate records are flagged and controlled for
4 – robust regressions
5 – weighted regressions

Notes: The duplicate records are randomly drawn from the overall distribution.
Bars show the share of coefficients' estimates (for all coefficients jointly) with an acceptable bias ($|Dfbetas| < 0.05$) for the five considered solutions (1 to 5).

1 doublet, from 47.2% to 61.8% with 6 doublets, from 28.4% to 41% with 16 doublets, from 16.3% to 25.7% with 40 doublets, and from 11.4% to 15.2% with 79 doublets.

Robust regression performs poorly in all cases. Figure 2 shows that robust regression may produce systematically biased coefficients: the median value of estimated coefficients is typically higher or lower than the unbiased estimates. Figure 3 shows that robust regression produces biased estimates more often than the 'naive' estimation, especially if records are duplicated multiple times (upper panel of Figure 3). The probability of obtaining unbiased coefficients is around 20-28% in most analyzed cases, and goes down to 12.2% when 79 doublets are included in the data.

3.3 'Typical' vs. 'deviant' cases

Tables 2 and 3 show the probability of obtaining unbiased estimates when the duplicated record was drawn from the center of the distribution (third column), from the lower tie (fourth column), and from the upper tie (fifth column) of the distribution of the y variable. For ease of comparison, the second column reports the chance of obtaining unbiased estimates for the 'unconstrained' case, i.e. when the duplicates were randomly drawn from the overall distribution (the same information reported in Figure 3). We expected that the risk of obtaining biased estimates is lower if the duplicate records are 'typical', i.e. they come from the center of the distribution, than if the duplicate records are 'deviant', i.e. they come from the upper of lower tie of the distribution.

Results in Table 2 show that the location of duplicate records in the distribution of the dependent variable makes little difference for the risk of the bias. The risk of bias in 'naive' estimation is higher for the ties than for the center for a doublet, triplet, and quintuplet, but not for a quadruplet and a sextuplet. Moreover, the risk of bias when the duplicate is drawn from the overall distribution is usually not lower than when the duplicate is drawn from the tie. These results do not depend on the solution adopted to deal with duplicates.

Inspection of Table 3 confirms that these results generally hold also for the cases included in Scenario 2. However, when large number of doublets is included in the data, the risk of bias is smaller when the duplicates are drawn from the overall distribution than when they are drawn from either the center or the ties of the distribution.

The same conclusions hold when the duplicated observations are randomly drawn on the basis of the distribution of the x rather than the y variable. Results are presented in Tables 6 and 7 in Appendix B.

Table 2: Probability of obtaining unbiased estimates for 'unconstrained', 'typical' and 'deviant' variants in scenario 1. Results refer to the following variants: when duplicates are randomly drawn from the overall, the center, the lower and the upper part of the distribution of the y variable.

	Duplicated observation drawn randomly from:			
	overall distribution	center of distribution	lower quartile	upper quartile
1 doublet:				
'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
1 triplet:				
'Naive' estimation	67.92	68.67	68.40	67.03
Drop duplicates	80.17	79.30	79.67	81.38
Flag and control	80.15	79.22	79.65	81.33
Robust regression	28.43	27.05	29.18	29.27
Weighted regression	79.80	79.03	79.47	80.80
1 quadruplet:				
'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
1 quintuplet:				
'Naive' estimation	46.27	48.08	46	47.35
Drop duplicates	66.50	68.67	68.22	69.08
Flag and control	66.40	68.63	68.15	69.03
Robust regression	25.90	25.13	27.30	25.23
Weighted regression	66.13	68.42	67.65	68.55
1 sextuplet:				
'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

Table 3: Probability of obtaining unbiased estimates for 'unconstrained', 'typical' and 'deviant' variants in scenario 2. Results refer to the following variants: when duplicates are randomly drawn from the overall, the center, the lower and the upper part of the distribution of the y variable.

	Duplicated observations drawn randomly from:			
	overall distribution	center of distribution	lower quartile	upper quartile
	1 doublet:			
'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
	6 doublets:			
'Naive' estimation	47.15	46.83	45	47.17
Drop duplicates	48.40	48.23	47.10	43.90
Flag and control	46.63	45.77	44.83	42.92
Robust regression	27.02	25.68	31.98	25.80
Weighted regression	61.80	63.60	62.05	62.83
	16 doublets:			
'Naive' estimation	28.38	29	27.38	28.93
Drop duplicates	29.20	27.88	27.05	29.68
Flag and control	28.57	29.40	26.85	25.43
Robust regression	21.35	21.43	26.38	20.48
Weighted regression	41.05	40.13	40.95	40.42
	40 doublets:			
'Naive' estimation	16.30	17.85	17.43	17.80
Drop duplicates	17.48	18.40	16.07	16.60
Flag and control	16.25	18.25	12.63	13.75
Robust regression	14.55	15.35	18.20	17.52
Weighted regression	25.70	25.63	24.25	25.65
	79 doublets:			
'Naive' estimation	11.43	12.07	10.25	12.53
Drop duplicates	11.95	11.48	9.575	12.28
Flag and control	11.70	12.55	6.500	7.400
Robust regression	12.18	10.03	10.98	11.63
Weighted regression	15.23	18.20	17.13	17.70

4 Conclusions

Availability of reliable data is a prerequisite for running appropriate analyses. This applies to any scientific discipline. In this paper we focused on the quality of survey data and, in particular, on the consequences of duplicate records for the reliability of regression estimates. Practitioners tend to take for granted the quality of the data at their disposal. This can explain the scarcity of papers dealing with the consequences of duplicate records for the reliability of regression estimates. Yet, a recent research by Slomczynski et al. (2015) raised awareness about the quality of survey data and warned about the possible consequences of ignoring the presence of duplicate records. Slomczynski et al. (2015) showed that a number of well-established and broadly used survey data are affected by duplicate records at various degrees. Unfortunately, little is known about the bias induced by duplicated observations in survey data.

Present paper fills this gap analyzing the effect of duplicate records on estimates obtained in OLS regression. We assessed the severity of the bias induced by duplicate records analyzing two scenarios: in the first one we focus on the bias induced by the number of duplications (from 1 to 5); in the second one we focus on the bias due to the number of duplicated records (from 1 to 79). Additionally, we assessed how the risk of obtaining biased estimates changes when the duplicates are situated in specific parts of data distribution (the center, the lower and the upper tie, or across the whole distribution). Finally, we compared the 'naive' estimation, which ignores the presence of duplicate records, with four alternative solutions to decrease the bias from the presence of duplicate records: excluding duplicates from the analysis; flagging duplicates and controlling for them in the estimation; using robust regression which weights down the impact of influential cases on the estimates; weighting the observations by the inverse of the duplicates' multiplicity.

To this aim we created an artificial dataset of $N = 1,500$ observations and four variables with a known covariance matrix. We adopted a Monte Carlo simulation with 1,000 replications to investigate the effect of 40 patterns of duplicate records on the bias of regression estimates. Furthermore, we used Dfbetas to assess the severity of the bias related to various patterns of duplicate records and to various solutions.

Results showed that the risk of obtaining biased estimates of regression coefficients increased with the number of duplicate records. If data contained a single sextuplet, i.e. less than 1% of the sample, the probability of obtaining unbiased estimates was 41.6%. If data contained 79 doublets of identical records, i.e. duplicates summed up to about 10% of the sample, the probability of obtaining unbiased estimates was about 11.4%.

Hence, even a small number of duplicate records created considerable risk of obtaining biased estimates. We emphasize that the patterns of duplicate records used in this analysis were consistent with those identified by Slomczynski et al. (2015), and they can, therefore, be regarded as realistic. This suggests that practitioners who fail to account for the presence of duplicate records in their analysis may reach misleading conclusions.

Additionally, our analysis provided evidence that the risk of bias was not lower if the duplicate records were located close to the center of the distribution of the dependent variable. The differences between 'typical', 'unconstrained' and 'deviant' variants were small. Even if duplicate observations were drawn from the center of a distribution, the risk of obtaining biased estimates remained high: 60.4% in case of a 'naive' regression run on data with a sextuplet of duplicated records, and 87.9% if data contain 79 doublets of duplicated records.

We also explored the effectiveness of five possible solutions to minimize the estimation bias induced by the presence of duplicated records. Weighting the duplicates by the inverse of their multiplicity was the best solution, among the considered ones, to minimize the estimation bias due to duplicated observations. This solution outperformed 'naive' estimates in presence of one doublet, and it performed equally to dropping or flagging the duplicates when one triplet, quadruplet, quintuplet or sextuplet were present in the data. Weighting by the inverse of the multiplicity was the best solution to minimize the bias also when the number of duplicated records increased. The performance of this solution decreased when the number of duplicates increased, but the chances of unbiased estimates were higher than in the alternative solutions.

Finally, robust regression, which weights down the impact of influential cases on the estimated regression coefficients, performed poorly in all cases. The risk of obtaining biased estimates was higher with robust regression than when duplicate cases are ignored in the analysis (‘naive’ estimation).

These results are discouraging, but not pessimistic: although duplicate data plague some of the major surveys currently used in social sciences, it is possible to adopt solutions to minimize the risk of biased estimates. Among the possible solutions available to practitioners, excluding duplicates from the analysis, flagging and controlling for the duplicated records, or weighting by the inverse of multiplicity seem reasonable solutions when a single observation has been duplicated, even multiple times. Yet, if multiple doublets are present in the data only weighting by the inverse of the multiplicity provides better results than a ‘naive’ estimation. This conclusion emphasizes the importance of obtaining data of high quality, because correcting the data with statistical tools is not a trivial task. This calls for further research about how to address the presence of multiple doublets in the data and more refined statistical tools to minimize the consequent estimation bias.

References

American Statistical Association (2003). Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects. Published in Survey Research Volume 35, Number 1, 2004, Newsletter from the Survey Research Laboratory, College of Urbaln Planning and Public Affairs, University of Illinois at Chicago.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16.

Kuriakose, N. and Robbins, M. (2015). Falsification in surveys: Detecting near duplicate observations. *Available at SSRN.* Accessed on 28th of July 2015.

Lessler, J. and Kalsbeek, W. (1992). *Nonsampling error in surveys.* Wiley, New York.

Slomczynski, K. M., Powałko, P., and Krauze, T. (2015). The large number of duplicate records in international survey projects: The need for data quality control. *CONSIRT Working Papers Series 8 at consirt.osu.edu.*

Waller, L. G. (2013). Interviewing the surveyors: Factors which contribute to questionnaire falsification (curbstoning) among Jamaican field surveyors. *International Journal of Social Research Methodology*, 16(2):155–164.

A Descriptive statistics for the simulated datasets.

Table 4: Descriptive statistics for the initial dataset and for exemplary simulated datasets in Scenario 1.

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

Table 5: Descriptive statistics for the initial dataset and for exemplary simulated datasets in Scenario 2.

N. of duplicates	variable	mean	sd	min	max	obs	missing
		3016	749.7	344.9	5775	1500	0
		6176	2899	−3213	17299	1500	0
Initial dataset		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
		3015	750.0	344.9	5775	1500	0
		6176	2899	−3213	17299	1500	0
1 doublet		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
		3015	750.2	344.9	5775	1500	0
		6174	2901	−3213	17299	1500	0
6 doublets		187.8	21.74	103.2	261.4	1500	0
		21.25	5.635	1.967	41.45	1500	0
	duplicates (flag)	0.00400	0.0631	0	1	1500	0
		3016	750.4	344.9	5775	1500	0
		6177	2885	−3213	17299	1500	0
16 doublets		187.8	21.77	103.2	261.4	1500	0
		21.26	5.625	1.967	41.45	1500	0
	duplicates (flag)	0.0107	0.103	0	1	1500	0
		3020	754.5	344.9	5775	1500	0
		6181	2897	−3213	17299	1500	0
40 doublets		187.9	21.82	103.2	261.4	1500	0
		21.23	5.654	1.967	41.45	1500	0
	duplicates (flag)	0.0267	0.161	0	1	1500	0
		3016	746.4	344.9	5775	1500	0
		6172	2886	−3213	17299	1500	0
79 doublets		187.9	21.58	103.2	261.4	1500	0
		21.27	5.589	1.967	41.45	1500	0
	duplicates (flag)	0.0527	0.223	0	1	1500	0

B Results when duplicates are drown from the distribution of x variable

Table 6: Probability of obtaining unbiased estimates for 'unconstrained', 'typical' and 'deviant' variants in scenario 1. Results refer to the following variants: when duplicates are randomly drawn from the overall, the center, the lower and the upper part of the distribution of the x variable.

	Duplicated observations drawn randomly from:			
	overall distribution	center of distribution	lower quartile	upper quartile
	1 doublet:			
'Naive' estimation	86.67	88.35	84.35	85.25
Drop duplicates	87	87.90	83.85	85.40
Flag and control	86.97	87.90	83.85	85.40
Robust regression	28.10	26.50	28.75	28.05
Weighted regression	94.10	93.55	93.97	93.78
	1 triplet:			
'Naive' estimation	67.92	72.38	66.30	63.45
Drop duplicates	80.17	81.28	79.55	79.10
Flag and control	80.15	81.25	79.53	79.08
Robust regression	28.43	26.98	28.20	28.38
Weighted regression	79.80	81.08	79.20	78.78
	1 quadruplet:			
'Naive' estimation	54.48	58.67	55.35	50.73
Drop duplicates	72.33	72.42	74.15	72.38
Flag and control	72.28	72.42	74.13	72.38
Robust regression	26.57	27.43	26.15	25.80
Weighted regression	71.92	72.08	73.75	71.83
	1 quintuplet:			
'Naive' estimation	46.27	46.38	46.05	44.10
Drop duplicates	66.50	69.67	66.92	67.13
Flag and control	66.40	69.63	66.88	67.05
Robust regression	25.90	24.55	24.02	24.05
Weighted regression	66.13	69.17	66.65	66.72
	1 sextuplet:			
'Naive' estimation	41.63	44.33	37.80	37.75
Drop duplicates	64.45	64.58	62.25	62.83
Flag and control	64.30	64.50	62.13	62.75
Robust regression	24.18	24.70	20.77	21.55
Weighted regression	63.90	64.03	61.75	62.52

Table 7: Probability of obtaining unbiased estimates for 'unconstrained', 'typical' and 'deviant' variants in scenario 2. Results refer to the following variants: when duplicates are randomly drawn from the overall, the center, the lower and the upper part of the distribution of the x variable.

	Duplicated observations drawn randomly from:			
	overall distribution	center of distribution	lower quartile	upper quartile
	1 doublet:			
'Naive' estimation	86.67	88.35	84.35	85.25
Drop duplicates	87	87.90	83.85	85.40
Flag and control	86.97	87.90	83.85	85.40
Robust regression	28.10	26.50	28.75	28.05
Weighted regression	94.10	93.55	93.97	93.78
	6 doublets:			
'Naive' estimation	45.98	49.38	43.73	44.33
Drop duplicates	45.92	47.80	45.88	45
Flag and control	46.40	47.50	44.70	44.88
Robust regression	27.48	28.13	28.35	26.30
Weighted regression	62.98	62.75	62.02	62.05
	16 doublets:			
'Naive' estimation	28.77	29.93	28.43	28.30
Drop duplicates	30.50	30.90	28.93	29.48
Flag and control	28.25	30.20	27.98	26.30
Robust regression	22.63	23.10	22.63	20.80
Weighted regression	39.42	39.30	40.77	39.60
	40 doublets:			
'Naive' estimation	17.75	18.05	17.20	18.13
Drop duplicates	18.18	18.80	16.65	16.48
Flag and control	17.85	18.27	16.25	18.30
Robust regression	16.82	16.40	16.80	14.63
Weighted regression	23.50	25.15	23.75	25.50
	79 doublets:			
'Naive' estimation	14.48	13.88	12.40	13.80
Drop duplicates	13.10	13.53	12.30	12.75
Flag and control	14.82	13.98	11.93	13.20
Robust regression	12.80	13.53	13.03	11.63
Weighted regression	18.05	18.18	16.43	17.88