92

Economie et Statistiques
Working papers du STATEC

march 2017

Bias and efficiency loss in regression estimates due to duplicated observations: a Monte Carlo simulation

Auteurs: Francesco Sarracino and Malgorzata Mikucka

Abstract

Recent studies documented that survey data contain duplicate records. We assess how duplicate records affect regression estimates, and we evaluate the effectiveness of solutions to deal with duplicate records. Results show that the chances of obtaining unbiased estimates when data contain 40 doublets (about 5% of the sample) range between 3.5% and 11.5% depending on the distribution of duplicates.

If 7 quintuplets are present in the data (2% of the sample), then the probability of obtaining biased estimates ranges between 11% and 20%. Weighting the duplicate records by the inverse of their multiplicity, or dropping superfluous duplicates outperform other solutions in all considered scenarios.

Our results illustrate the risk of using data in presence of duplicate records and call for further research on strategies to analyse affected data.

Keywords: duplicated observations, estimation bias, Monte Carlo simulation, inference.

Bias and efficiency loss in regression estimates due to duplicated observations: a Monte Carlo simulation*

Francesco Sarracino† and Małgorzata Mikucka‡

February 3, 2017

Abstract

Recent studies documented that survey data contain duplicate records. We assess how duplicate records affect regression estimates, and we evaluate the effectiveness of solutions to deal with duplicate records. Results show that the chances of obtaining unbiased estimates when data contain 40 doublets (about 5% of the sample) range between 3.5% and 11.5% depending on the distribution of duplicates. If 7 quintuplets are present in the data (2% of the sample), then the probability of obtaining biased estimates ranges between 11% and 20%. Weighting the duplicate records by the inverse of their multiplicity, or dropping superfluous duplicates outperform other solutions in all considered scenarios. Our results illustrate the risk of using data in presence of duplicate records and call for further research on strategies to analyze affected data.

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

1 Introduction

To achieve reliable results survey data must accurately report respondents' answers. Yet, sometimes they don't. A recent study by Slomczynski et al. (2015) investigated survey projects 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, and it is unlikely that two respondents provide identical answers to all (or nearly all) substantive survey questions (Hill, 1999). 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 fakes. 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; Diekmann, 2005; Koczela et al., 2015; Kuriakose and Robbins, 2015; Waller, 2013).

*This article reflects the view of the authors and do not engage in any way STATEC research, ANEC and funding partners. The authors gratefully acknowledge the support of the Observatoire de la Compétitivité, Ministère de l'Economie, DG Compétitivité, Luxembourg and STATEC. 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.

†STATEC Research, Institut national de la statistique et des études économiques du Grand-Duché du Luxembourg, ANEC, 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

‡MZES, Mannheim University (Germany), Université catholique de Louvain (Belgium) and LCSR National Research University Higher School of Economics, Russian Federation.

1.1 Duplicate records in social survey data

Slomczynski et al. (2015) analyzed 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 duplicate 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, different numbers of countries were affected. Latinobarometro is an extreme case where surveys from 13 out of 19 countries contained duplicate 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 records.

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 duplicate records. However, in 48% of surveys containing duplicates they found various patterns of duplicate records, such as multiple doublets (i.e. multiple pairs of identical records) or identical records repeated three, four, or more times. For instance, the authors identified 733 duplicate 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 duplicate records consisted of 27 doublets, 36 duplicate 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 duplicate records, i.e. 11.0% of the sample).

These figures refer to full duplicates. However, other research analyzed the prevalence of near duplicates, that is records which differ for only a small number of variables. Kuriakose and Robbins (2015) analyzed near duplicates in data sets commonly used in social sciences and showed that 16% of analyzed surveys revealed a high risk of widespread falsification with near duplicates. The authors emphasized that demographic and geographical variables are rarely falsified, because they typically need to meet the sampling frame. Behavioral and attitudinal variables, on the other hand, were falsified more often. In such cases, interviewers may copy only selected sequences of answers from other respondents, so that the correlations between variables are as expected, and the forgery remains undetected.

1.2 Implications for estimation results

Duplicate records may affect statistical inference in various ways. If duplicate records introduce 'random noise', then they may produce an attenuation bias, i.e. bias the estimated coefficient towards zero (Finn and Ranchhod, 2013). However, if the duplicate records do not introduce random noise, they may bias the estimated correlations in other directions. The size of the bias should increase with the number of duplicate interviews, and it should depend on the difference of covariances and averages between the original and duplicate interviews (Schräpler and Wagner, 2005).

On the other hand, duplicate records can reduce the variance, and thus they may artificially increase the statistical power of estimation techniques. The result is the opposite of the attenuation bias: narrower estimated confidence intervals and stronger estimated relationships among variables (Kuriakose and Robbins, 2015). In turn, this may increase the statistical significance of the coefficients and affect the substantive conclusions.

The risk of obtaining biased estimates may differ according to the characteristics of the observations being duplicated. Slomczynski et al. (2015) suggested that 'typical' cases, i.e. the duplicate records located near the median of a variable, may affect estimates less than 'deviant' cases, i.e. duplicate records located close to the ties of the distribution.

The literature on how duplicate records affect estimates from regression analysis, and how to deal with them is virtually not existing. Past studies focused mainly on strategies to identify duplicate and near-

duplicate records (Elmagarmid et al., 2007; Hassanzadeh and Miller, 2009; Kuriakose and Robbins, 2015; Schreiner et al., 1988). However some studies analyzed how intentionally falsified interviews (other than duplicates) affected summary statistics and estimation results. Schnell (1991) studied the consequences of including purposefully falsified interviews in the 1988 German General Social Survey (ALLBUS). The results showed a negligible impact on the mean and standard deviation. However, the falsified responses produced stronger correlations between objective and subjective measures, more consistent scales (with higher Cronbach's α), higher R^2, and more significant predictors in OLS regression. More recently, Schräpler and Wagner (2005) and Finn and Ranchhod (2013) did not confirm the greater consistency of falsified data. On the contrary, they showed a negligible effect of falsified interviews on estimation bias and efficiency.

1.3 Current analysis

This is the first analysis assessing how duplicate records affect the bias and efficiency of regression estimates. We focus on two research questions: first, how do duplicates affect regression estimates? Second, how effective are the possible solutions? We use a Monte Carlo simulation, a technique for generating random samples on a computer to study the consequences of probabilistic events (Ferrarini, 2011; Fishman, 2005). In our simulations we consider three scenarios of duplicate data:

Scenario 1. when one record is multiplied several times (a sextuplet, an octuplet, and a decuplet),

Scenario 2. when several records are duplicated once (16, 40, and 79 doublets, which correspond to 2%, 5% and 10% of the sample respectively),

Scenario 3. when several records are duplicated four times (7, 16, and 31 quintuplets, which correspond to 2%, 5% and 10% of the sample respectively).

We chose the number of duplicates to mimic the results provided by Slomczynski et al. (2015). We also investigate how regression estimates change when duplicates are located in specific parts of the distribution of the dependent variable. We evaluate four variants, namely:

Variant i. 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);

Variant ii. when they are chosen randomly between the first and third quartile of the dependent variable (i.e. when they are located around the median: this is the 'typical' variant);

Variant iii. when they are chosen randomly below the first quartile of the dependent variable (this is the first 'deviant' variant);

Variant iv. when they are chosen randomly above the third quartile of the dependent variable (this is the second 'deviant' variant).

We expect that Variants *iii* and *iv* affect regression estimates more than Variant *i*, and that Variant *ii* affects them the least. Additionally, to check the robustness of our findings, we replicate the whole analysis to check how the position on the distribution of one of the independent variables affects regression estimates.

For each scenario and variant we compute the following measures to assess how duplicates affect regression estimates:

Measure A. standardized bias;

Measure B. risk of obtaining biased estimates, as measured by Dfbetas;

Measure C. Root Mean Square Error (RMSE), which informs about the efficiency of the estimates.

We consider five solutions to deal with duplicate records, and we assess their ability to reduce the bias and the efficiency loss:

Solution a. 'naive estimation', i.e. analyzing the data as if they were correct;

Solution b. dropping all the duplicates from the data;

Solution c. flagging the duplicate records and including the flag among the predictors;

Solution d. dropping all superfluous duplicates;

Solution e. weighting the duplicate records by the inverse of their multiplicity.

Finally, we check the sensitivity of our results to the sample size. Our basic analysis uses a sample of $N = 1,500$, because many nationally representative surveys provide samples of similar sizes. However, we also run the simulation for samples of $N = 500$ and $N = 5,000$ to check the robustness of our results to the chosen sample sizes.

2 Method

To assess how duplicate records affect the results of OLS regression we use a Monte Carlo simulation (Ferrarini, 2011; Fishman, 2005). The reason is that we need an artificial data set where the relationship among variables is known, and in which we iteratively manipulate the number and distribution of duplicates. The random element in our simulation is the choice of records to be duplicated, and the choice of observations which are replaced by duplicates. At each iteration, we compare the regression coefficients in presence of duplicates with the true coefficients to tell whether duplicates affect regression estimates.

Our analysis consists of four steps. First, we generate the initial data set. Second, we duplicate randomly selected observations according to the three 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 (Solution a); we also estimate regression models using the four alternative solutions $(b - e)$ to deal with duplicate records. Finally, we compute the standardized bias, the risk of obtaining biased estimates, and the Root Mean Square Error. Figure 1 summarizes our strategy.

2.1 Data generation

We begin by generating a data set of $N = 1,500$ observations which contains three variables: x, z, and t. We create the original data set using random normally distributed variables with a known correlation matrix (shown in Table 1). The correlation matrix is meant to mimic real survey data and it is based on the correlation of household income, age, and number of hours worked as retrieved from the sixth wave of the European Social Survey (2015).

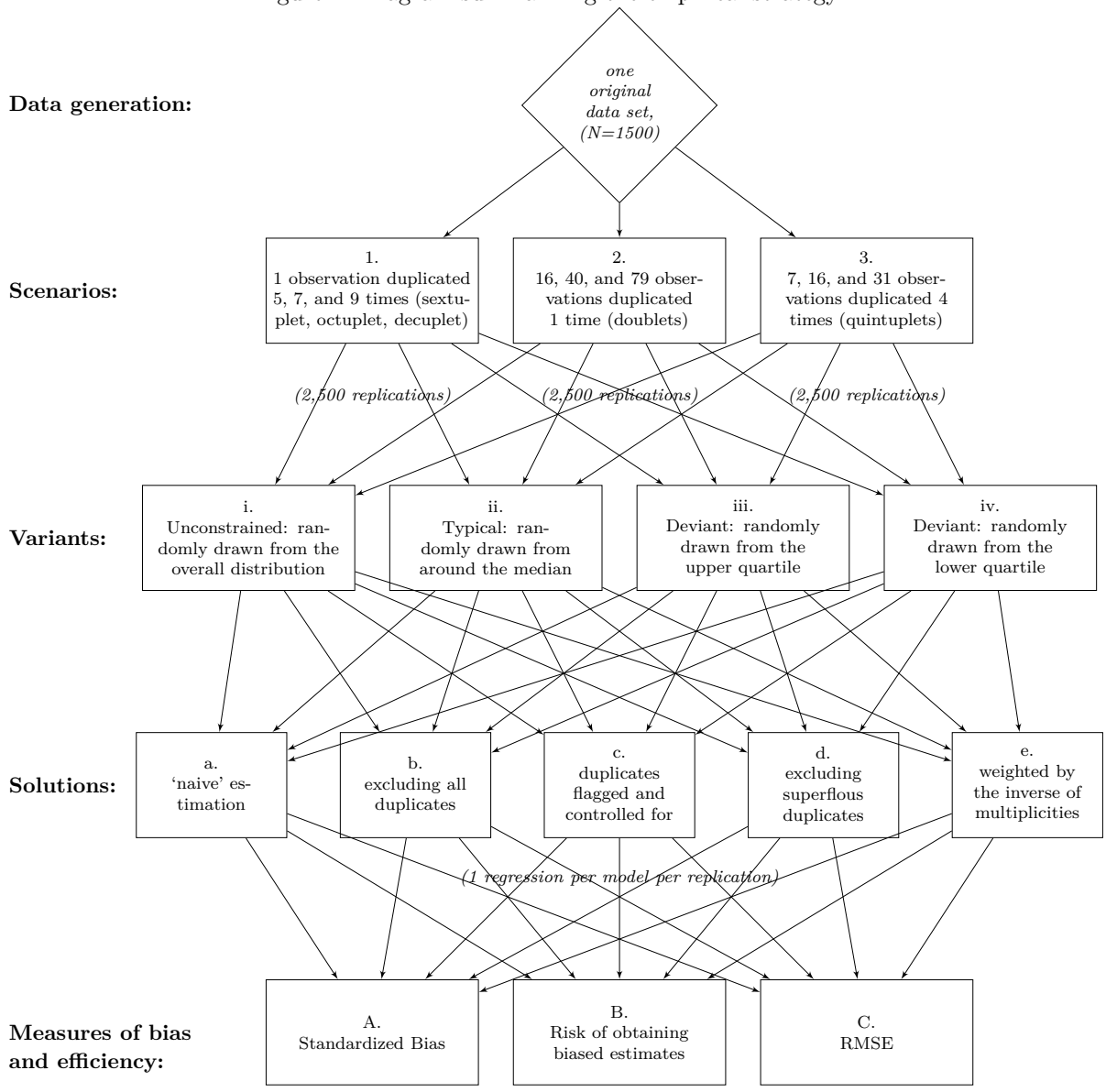
Table 1: Matrix of correlations used to generate the original data set.

variables	x	z	t
x	1		
z	−0.04	1	
t	0.09	−0.06	1

We generate the dependent variable (y) as a linear function of x, z, and t as reported in Equation 1:

$$y_i = 5.36 - 0.04 \cdot x_i + 0.16 \cdot z_i + 0.023 \cdot t_i + \epsilon_i \tag{1}$$

Figure 1: Diagram summarizing the empirical strategy.



where the coefficients are retrieved from the sixth wave of the European Social Survey. All variables and the error term ϵ_i are normally distributed. The descriptive statistics of the generated data set are shown in the first four lines of Table 5 in Appendix A.

2.2 Duplicating selected observations

In the second step we use a Monte Carlo simulation to generate duplicate records, which replace randomly chosen original records, i.e. the interviews that would have been conducted if no duplicates had been introduced in the data. This strategy is motivated by the assumption that duplicate records substitute for authentic interviews. Thus, if duplicates are present, researchers do not only face the risk of fake or erroneous information, but they also lose information from genuine respondents. We duplicate selected observations in three scenarios (each comprising three cases) and in four variants. Thus, overall we investigate 36 patterns ($3 \cdot 3 \cdot 4 = 36$) of duplicate records. For each pattern we run 2,500 replications.

Scenario 1: a sextuplet, an octuplet, and a decuplet In the first scenario we duplicate one randomly chosen record 5, 7, and 9 times, thus introducing in the data a sextuplet, an octuplet, and a decuplet of identical observations which replace for 5, 7, and 9 original observations. These cases are possible in the light of the analysis by Slomczynski et al. (2015) who identified in real survey data instances of octuplets. In this scenario the share of duplicates in the sample is very small, ranging from 0,40% for a sextuplet to 0,67% for a decuplet.

Scenario 2: 16, 40, and 79 doublets In the second scenario we duplicate sets of 16, 40, and 79 randomly chosen observations one time, creating 16, 40, and 79 pairs of identical observations (doublets). In this scenario the share of duplicates is 2,13% (16 doublets), 5,33% (40 doublets), and 10,53% (79 doublets). These shares are consistent with the results by Slomczynski et al. (2015), as in their analysis about 15% of the affected surveys had 10% or more duplicate records.

Scenario 3: 7, 16, and 31 quintuplets In the third scenario we duplicate sets of 7, 16 and 31 randomly chosen observations 4 times, creating 7, 16 and 31 quintuplets. They replace, for 28, 64, and 124 true records respectively. In this scenario the share of duplicate records is 2,33% (7 quintuplets), 5,33% (16 quintuplets), and 10,33% (31 quintuplets).

To check whether the position of duplicates in the distribution matters, we run each of the scenarios in four variants, as presented in Figure 2.

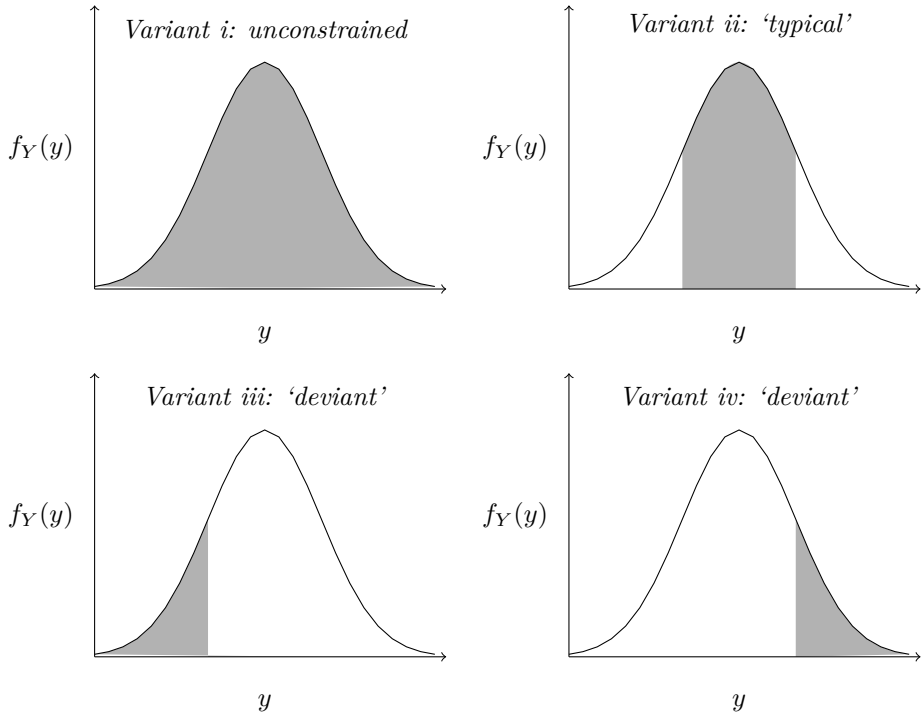
Variant *i* ('unconstrained'): the duplicates and the replaced interviews are randomly drawn from the overall distribution of the dependent variable.

Variant *ii* ('typical'): the duplicates are randomly drawn from the values around the median of the dependent variable, i.e. between the first and third quartile, and the replaced interviews are drawn from the overall distribution.

Variant *iii* and *iv* ('deviant'): In Variant *iii* the duplicates are randomly drawn from the lower quartile of the dependent variable; in Variant *iv* they are randomly drawn from the upper quartile of the dependent variable. The replaced interviews are drawn from the overall distribution.

To illustrate our data, Table 5 in Appendix A reports the descriptive statistics of some of the data sets produced during the replications (lines 5 to 45).

Figure 2: Presentation of the Variants *i-iv* used in the Monte Carlo simulation.



2.3 'Naive' estimation and possible solutions

In the third step we run a 'naive' estimation which takes data as they are, and subsequently we investigate the remaining four alternative solutions to deal with duplicates. For each solution we estimate the following model:

$$y_i = \alpha + \beta_x \cdot x_i + \beta_z \cdot z_i + \beta_t \cdot t_i + \varepsilon_i \tag{2}$$

Solution *a*: '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 standardized bias, the risk of obtaining biased estimates, and the Root Mean Square Error resulting from the mere presence of duplicate records (see Section 2.4).

Solution *b*: Drop all duplicates In Solution *b* we drop all duplicates, including the observations that may come from true interviews. We consider such a case, because, if records are identical on some, but not on all variables (most likely, differences may exist on demographic and geographical variables to reflect the sampling scheme), then it is not obvious to tell the original observations from the fake duplicates. It is also possible that all duplicates are forged and should be excluded from the data. Therefore, rather that deleting the superfluous duplicates and retaining the original records, we exclude all duplicate records from the data at the cost of reducing the sample size.

Solution *c*: Flag duplicated observations and control for them This solution is similar to the previous one because we identify all duplicate records as suspicious. However, rather than dropping them, we generate a dichotomous variable (duplicate = 1, otherwise = 0), and include it among the predictors in Equation 2. This solution was suggested by Slomczynski et al. (2015) as a way to control for the error generated by duplicate records.

Solution *d*: Drop superfluous duplicates "[E]liminating duplicate and near duplicate observations from analysis is imperative to ensuring valid inferences" (Kuriakose and Robbins, 2015, p. 2). Hence, we

delete superfluous duplicates to retain a sample of unique records. The difference compared to Solution *b* is that we keep one record for each set of duplicates and we drop all the superfluous records.

Solution *e*: Weight by the inverse of multiplicities This method has been proposed by Lessler and Kalsbeek (1992). We construct a weight which takes the value of 1 for unique records, and the value of the inverse of multiplicity for duplicate records. For example, the weight takes the value 0.5 for doublets, 0.2 for quintuplets, 0.1 for decuplets, etc. Subsequently, we use these weights to estimate Equation 2.

2.4 The assessment of bias and efficiency

We use three measures to assess the consequences of duplicates for regression estimates, and to investigate the efficiency of the alternative solutions to deal with them.

Measure A: Standardized bias This macro measure of bias informs whether a coefficient is systematically over or under estimated. It is computed as follows:

$$\text{Bias} = \left(\frac{\overline{\widehat{\beta}} - \beta}{SE(\widehat{\beta}_i)} \right) \cdot 100 \tag{3}$$

where β is the true coefficient from Equation 1, $SE(\widehat{\beta}_i)$ equals the standard deviation of empirically estimated coefficients ($\widehat{\beta}_i$), and $\overline{\widehat{\beta}}$ is the average of $\widehat{\beta}_i$.

Measure B: Risk of obtaining biased estimates To assess the risk of obtaining biased estimates in a specific replication, we resort to Dfbetas, which are normalized measures of how much specific observations (in our case the duplicates) affect the estimates of regression coefficients. Dfbetas are defined as the difference between the estimated and the true coefficients, expressed in relation to the standard error of the true estimate (see Equation 4).

$$\text{Dfbeta}_i = \frac{\widehat{\beta}_i - \beta}{SE(\widehat{\beta}_i)} \tag{4}$$

where i indicates the replication. Dfbetas measure the bias of a specific estimation. Thus, they complement standardized bias by informing about the risk of obtaining biased estimates. The risk is computed according to Equation 5. We set the cutoff value to 0.5, i.e. we consider the estimation as biased if the coefficients differ from the true values by more than half standard deviation.[1]

$$dummy_i = \begin{cases} 1 & \text{if } |\text{Dfbeta}_i| > 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$$Pr(bias) = \left(\frac{\sum_{i=1}^{n=2500} dummy_i}{n}\right) \cdot 100$$

Measure C: Root Mean Square Error (RMSE) Root mean square error is an overall measure of quality of the prediction and reflects both its bias and its dispersion (see Equation 6).

$$RMSE = \sqrt{\left(\overline{\widehat{\beta}} - \beta\right)^2 + \left(SE(\widehat{\beta}_i)\right)^2} \tag{6}$$

The RMSE is expressed in the same units of the variables (in this case the β coefficients), and it has no clear-cut threshold value. For ease of interpretation we express RMSE as a percentage of the respective coefficients from Equation 1, thus we report the normalized RMSE.

[1] This cutoff value is more conservative than the customary assumed value of $\frac{2}{\sqrt{N}}$, which, for $N = 1,500$ leads to the threshold value of 0.05.

3 Results

3.1 Standardized bias

Table 2 shows the standardized bias of the coefficient β_x for the considered scenarios, variants, and solutions. The results for the other coefficients are reported in Tables 6 – 8 in Appendix B. The third column of Table 2 contains information about the standardized bias for solution a, i.e. the 'naive estimation'.

Table 2: Standardized bias of the β_x coefficient (as a percentage of standard error).

Variant		Solution:				
		a 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
			Scenario 1			
Variant i: *'unconstrained'*	1 sextuplet	1.6	−2.6	−2.6	−1.7	−1.7
	1 octuplet	−0.3	0.7	0.7	0.6	0.6
	1 decuplet	−0.1	2.1	2.1	1.9	1.9
Variant ii: *'typical'*	1 sextuplet	32.0	−10.1	−10.1	−0.2	−0.2
	1 octuplet	31.4	−10.9	−10.9	−2.8	−2.8
	1 decuplet	36.3	−8.1	−8.1	−0.5	−0.5
Variant iii: *'deviant'*	1 sextuplet	−19.5	10.1	10.1	0.4	0.4
	1 octuplet	−18.6	4.5	4.5	−3.6	−3.6
	1 decuplet	−22.1	3.7	3.7	−4.7	−4.7
Variant iv: *'deviant'*	1 sextuplet	−14.8	7.9	7.9	0.4	0.4
	1 octuplet	−13.2	2.3	2.3	−3.5	−3.5
	1 decuplet	−10.0	−0.5	−0.5	−4.5	−4.5
			Scenario 2			
Variant i: *'unconstrained'*	16 doublets	2.9	2.6	2.9	3.9	3.9
	40 doublets	−1.6	−1.5	−1.4	−2.1	−2.1
	79 doublets	−0.9	−2.3	−0.8	−2.1	−2.1
Variant ii: *'typical'*	16 doublets	78.4	−78.1	68.6	1.3	1.3
	40 doublets	125.1	−116.1	118.9	4.5	4.5
	79 doublets	177.3	−165.6	172.3	9.3	9.3
Variant iii: *'deviant'*	16 doublets	−63.3	65.9	193.1	−1.7	−1.7
	40 doublets	−102.0	109.8	315.6	−3.2	−3.2
	79 doublets	−148.8	164.0	472.1	−14.3	−14.3
Variant iv: *'deviant'*	16 doublets	−42.6	43.0	137.1	−0.7	−0.7
	40 doublets	−65.9	74.5	245.1	−2.5	−2.5
	79 doublets	−92.2	114.1	359.6	−4.0	−4.0
			Scenario 3			
Variant i: *'unconstrained'*	7 quintuplets	−2.5	4.3	−3.1	2.7	2.7
	16 quintuplets	−1.0	2.4	−1.2	1.8	1.8
	31 quintuplets	0.4	0.6	−0.0	0.9	0.9
Variant ii: *'typical'*	7 quintuplets	81.6	−27.2	70.1	1.1	1.1
	16 quintuplets	126.1	−38.8	117.3	4.7	4.7
	31 quintuplets	172.0	−56.4	167.2	4.7	4.7
Variant iii: *'deviant'*	7 quintuplets	−48.2	28.0	110.0	−0.3	−0.3
	16 quintuplets	−69.1	43.5	172.3	−1.2	−1.2
	31 quintuplets	−95.6	60.3	246.4	−6.1	−6.1
Variant iv: *'deviant'*	7 quintuplets	−33.1	19.9	91.2	−0.9	−0.9
	16 quintuplets	−47.2	29.7	145.0	−0.8	−0.8
	31 quintuplets	−65.0	41.0	213.8	−6.5	−6.5

Overall, the standardized bias varies between nearly zero (1 decuplet in Variant i) to about 1,75 standard error for 79 doublets (177%) and 31 quintuplets (172%) in Variant ii, in Scenarios 2 and 3 respectively. The maximum bias for β_z is 120%, for β_t it is 72%, and for the intercept it is 120% (see Appendix B). In each scenario, the bias is highest for Variant ii (when the duplicates are located in the center of the distribution), and the smallest in Variant i (when the duplicates are randomly distributed). However, for the intercept the bias is highest when the duplicates are located on the ties (see Table 8

in Appendix B). Furthermore, the standardized bias increases with the share of duplicates in the data, but the composition of duplicates makes little difference. Duplicates consisting of quintuplets produce marginally smaller bias than duplicates consisting of doublets.

Among the four alternative solutions to deal with duplicates, solutions *d* and *e*, i.e. dropping the superfluous duplicates and weighting by the inverse of multiplicity perform best in all three scenarios, where they produce equivalent results. They reduce the standardized bias to less than 15%, and in most cases to less than 5%. On the other hand, dropping all duplicates (solution *b*) reduces the bias to a lesser extent, mainly in Scenarios 1 and 3, whereas in some cases of Scenario 2 this solution increases the bias. Flagging duplicates and controlling for them in the regression (solution *c*) performs worst, systematically increasing the bias in Scenarios 2 and 3, up to the value of almost 5 standard errors (β_x, Scenario 2, 79 doublets, Variant iii.).

In sum, duplicates can produce bias in regression estimates, especially when they are concentrated in the middle or on the ties of the distribution: the higher the number of duplicates, the stronger the bias. Dropping superfluous duplicates or weighting by the inverse of their multiplicity are effective ways to reduce the bias.

3.2 Risk of obtaining biased estimates

While standardized bias informs about the average bias due to duplicates, it is plausible that replications have upward and downward biases which, on average, offset each other. In other words, even with moderate standardized bias researchers risk to obtain biased estimates in specific estimations. To address this issue we turn to the analysis of Dfbetas.

Figure 3 shows box and whiskers diagrams of Dfbetas in Scenarios 2 (upper panel) and 3 (lower panel) for Variant *i*, i.e. when the duplicates are randomly drawn from the overall distribution of the dependent variable. We do not report results for Scenario 1 because in this case the risk of obtaining biased estimates is virtually zero. Results, however, are detailed in Table 3. On the *y*-axis we report the Dfbetas, on the *x*-axis we report the coefficients and the solutions. The two horizontal solid lines identify the cutoff values of Dfbetas (0.5) separating the replications with acceptable bias from the unacceptable ones. The diagrams show that the range of Dfbetas increases with the share of duplicates in the data (from 2 to 10%). The range is larger for quintuplets (Scenario 3) than for doublets (Scenario 2).

The average probability of obtaining unbiased estimates for all coefficients is shown in Table 3. Column *a* shows that, in case of 'naive' estimations, the risk of biased estimates varies from 0,14% (Scenario 1, one sextuplet, Variant *ii*) to about 58% (Scenario 3, 31 quintuplets, Variants *iii* and *iv*). In Scenario 1 the risk is small, with 89%-99% probability of obtaining unbiased estimates.

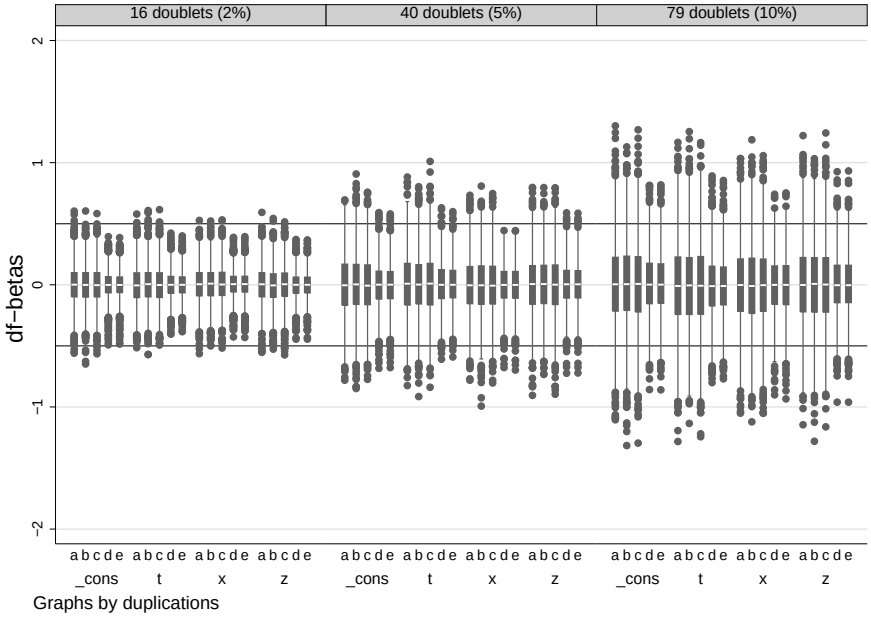
The results show three regularities. First, the risk of obtaining biased estimates increases with the share of duplicates in the data: it is $0.19\% - 0.70\%$ (depending on the variant) for 16 doublets, but it grows to $14.00\% - 32.48\%$ when 79 doublets are included in the data. In Scenario 3 it grows from $2.99\% - 20.19\%$ for 7 quintuplets, to $42.95\% - 58.62\%$ when data contain 31 quintuplets.

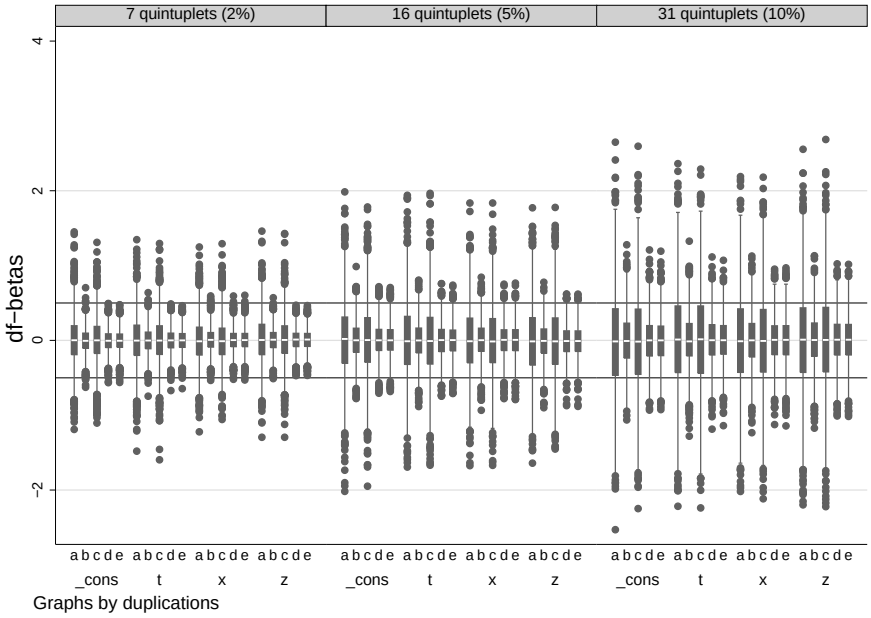
Second, when the duplicates constitute the same share of the data, the risk of obtaining biased estimates is higher for quintuplets than for doublets. For example, when duplicates constitute 2% of the sample, the risk of obtaining biased estimates is below 1% if they are 16 doublets, but ranges between 3% and 20% (depending on the variant) for 7 quintuplets. When duplicates constitute 10% of the data, the risk of obtaining biased estimates is $14\% - 32\%$ in case of 79 doublets, but $43\% - 58\%$ for 31 quintuplets.

Third, the risk of obtaining biased estimates is the highest in Variants *iii* and *iv*, i.e. when the duplicates are located on the ties, and lowest in Variant *ii*, when the duplicates are located around the median. For example, with 7 quintuplets, the probability of obtaining biased estimates is about 3% in Variant *ii*, but rises to about 20% in Variants *iii* and *iv*. For 31 quintuplets the risk is about 43% in Variant *ii*, but over 58% in Variants *iii* and *iv*.

As in case of standardized bias, weighting by the inverse of the multiplicity (Solution *e*), and dropping

Figure 3: Box and whiskers diagrams of Dfbetas in Scenario 2 and 3, Variant i.





a – 'Naive' estimation; b – Drop all duplicates; c – Flag and control; d – Drop superfluous duplicates; e – Weighted regression.

Notes: The duplicate records are randomly drawn from the overall distribution.
Box and whiskers show the distribution of Dfbetas (across 2,500 replications) for each of the coefficients in the model (the constant (*_cons*), β_x (x) , β_z (z) and β_t (t)) and for the solutions a to e.

Table 3: Probability of obtaining unbiased estimates ($\text{Dfbetas}_i < 0.5$).

Variant		Solution:				
		a 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
		Scenario 1				
Variant i: *'unconstrained'*	1 sextuplet	99.19	100	100	100	100
	1 octuplet	97.03	100	100	100	100
	1 decuplet	94.43	99.99	99.99	99.99	99.99
Variant ii: *'typical'*	1 sextuplet	99.86	100	100	100	100
	1 octuplet	99.66	100	100	100	100
	1 decuplet	98.90	100	100	100	100
Variant iii: *'deviant'*	1 sextuplet	97.90	100	100	100	100
	1 octuplet	94.66	100	100	100	100
	1 decuplet	90.89	100	100	100	100
Variant iv: *'deviant'*	1 sextuplet	98.10	100	100	100	100
	1 octuplet	94.47	100	100	100	100
	1 decuplet	89.18	100	100	100	100
		Scenario 2				
Variant i: *'unconstrained'*	16 doublets	99.81	99.82	99.82	100.00	100.00
	40 doublets	96.52	96.35	96.52	99.57	99.65
	79 doublets	86.00	86.90	86.07	96.13	96.37
Variant ii: *'typical'*	16 doublets	99.97	99.96	99.97	100.00	100.00
	40 doublets	96.29	96.80	96.56	99.62	99.70
	79 doublets	77.12	80.52	78.04	96.11	96.46
Variant iii: *'deviant'*	16 doublets	99.33	99.18	96.64	99.99	99.99
	40 doublets	88.52	86.77	68.29	99.63	99.67
	79 doublets	66.52	60.71	44.11	96.80	96.85
Variant iv: *'deviant'*	16 doublets	99.30	98.98	97.44	100.00	100.00
	40 doublets	89.20	87.10	63.65	99.64	99.67
	79 doublets	67.70	61.23	25.93	96.77	96.90
		Scenario 3				
Variant i: *'unconstrained'*	7 quintuplets	89.17	99.73	91.00	99.93	99.92
	16 quintuplets	71.89	96.38	72.36	97.99	98.17
	31 quintuplets	55.72	87.31	56.38	91.20	91.78
Variant ii: *'typical'*	7 quintuplets	97.01	99.78	98.00	99.86	99.88
	16 quintuplets	80.69	96.78	82.31	97.70	98.12
	31 quintuplets	57.05	87.81	58.45	90.63	91.40
Variant iii: *'deviant'*	7 quintuplets	80.33	99.60	95.30	99.96	99.98
	16 quintuplets	58.21	93.85	72.15	97.75	98.06
	31 quintuplets	41.38	81.46	48.47	90.69	91.03
Variant iv: *'deviant'*	7 quintuplets	79.91	99.71	95.20	99.97	99.97
	16 quintuplets	57.14	94.13	72.40	97.92	98.12
	31 quintuplets	41.88	81.67	44.21	91.41	91.74

the superfluous duplicates (Solution d) perform better than other solutions (see Table 3 and Figure 3). In Scenario 2, when doublets constitute about 5% of the data, these two solutions reduce the probability of obtaining biased estimates from about 4% (in Variants i and ii) or about 11% (Variants iii and iv) to under 1% in all cases. In case of 79 doublets, i.e when duplicates constitute about 10% of the sample, the risk of obtaining biased estimates reduces to about 3%, independently from the location of the duplicates, whereas it ranges between 14% and 33% for the naive estimation. In Scenario 3, when quintuplets constitute about 5% of the data, the risk of obtaining biased estimates declines from $19-43\%$ (depending on the variant) to about 2%. When quintuplets constitute about 10% of the sample, solutions d and e decrease the risk of obtaining biased estimates from $43-59\%$ to about 9%.

To sum up, weighting by the inverse of multiplicity and dropping the superfluous duplicates are the most effective solutions among the examined ones. Moreover, they perform particularly well when the duplicates are located on the ties of the distribution, i.e. when the risk of bias is the highest.

On the other hand, Solutions b (excluding all duplicates) and c (flagging the duplicates) perform worse. Flagging duplicates and controlling for them (c) fails to reduce the risk of obtaining biased estimates in both Scenarios 2 and 3. Excluding all duplicates (b) reduces the risk of obtaining biased estimates in Scenario 3 (quintuplets), but it performs poorly in Scenario 2: if the doublets are located on the ties, then flagging and controlling for them decreases the probability of obtaining unbiased estimates.

3.3 Root Mean Square Error

Table 4 shows the values of normalized RMSE for coefficient β_x. Results for other coefficients are available in Tables 9 – 11 in Appendix C. The loss of efficiency due to duplicates is overall small, reaching maximum 9% of the β_x coefficient (31 quintuplets in Variant iii), 13% of β_z, 20% of β_t, and 6% of the intercept. Consistently with previous results, solutions d and e, i.e. dropping superfluous duplicates and weighting the data perform best. In contrast to that, flagging the duplicates and controlling for them (solution c) performs poorly, and further reduces the efficiency of the estimates.

3.4 Robustness

3.4.1 Varying sample size

By setting up our experiment, we arbitrarily chose a sample size of $N = 1,500$ observations to mimic the average size of many of the publicly available social surveys. To check whether our results are independent from our choice, we repeated the experiment using two alternative samples: $N = 500$ and $N = 5,000$. In Figure 4 we report the results for Scenario 2, Variant i. The complete set of results is available upon request.

Figure 4 shows that the dispersion of the estimates with respect to the true values increases when the number of duplicates increases. Neglecting the presence of duplicates creates some problems when the share of duplicates reaches about 5% of the sample.

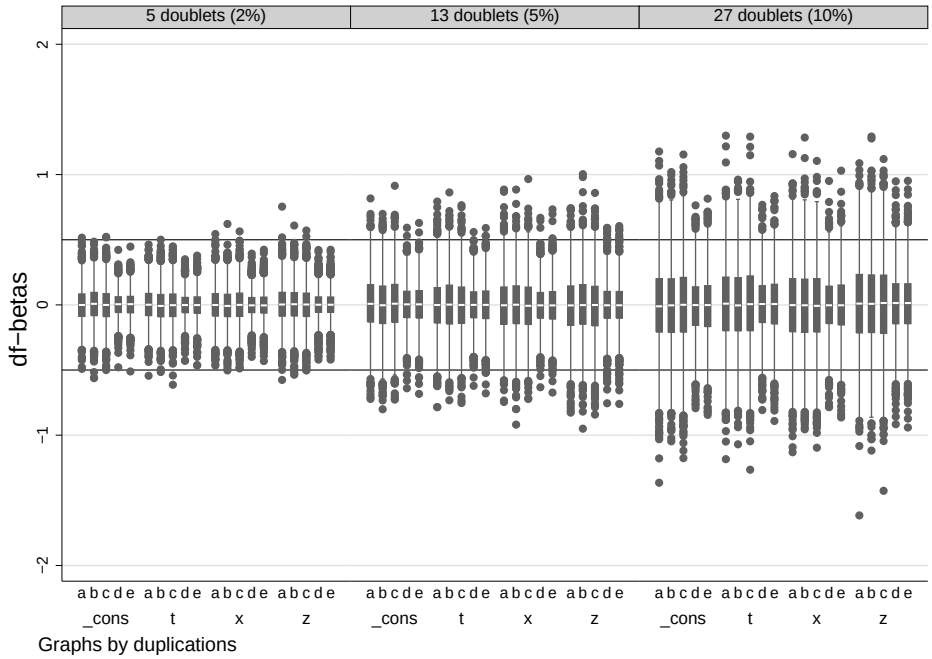
For $N = 500$ the probabilities of obtaining biased estimates amount to 2% and 11.4% when the doublets constitute 5% and 10% of the sample respectively. For $N = 5,000$ the same probabilities are 3% and 13%. These values are fairly in line with the results obtained for $N = 1,500$ for Variant i (3.5% and 14%).

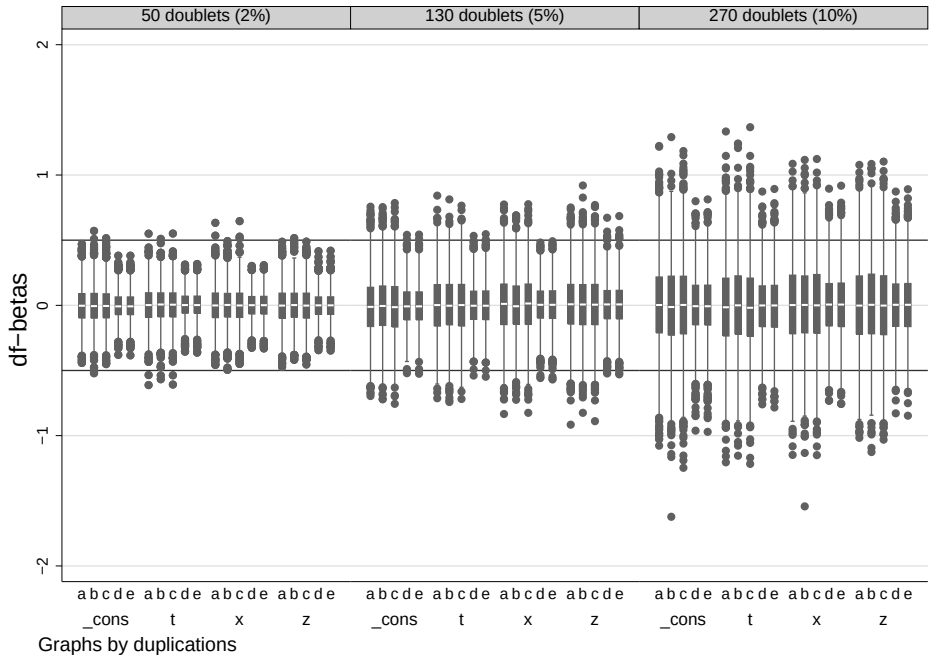
Consistently with the results for $N = 1,500$, weighting by the inverse of the multiplicity or dropping all superfluous duplicates are most effective in reducing the risk of obtaining biased estimates. Our conclusion about the influence of duplicated records and the efficiency of the solutions does not depend on sample size.

Table 4: Normalized RMSE of the β_x coefficient (in percentage).

Variant		Solution:				
		a 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
		Scenario 1				
Variant i: 'unconstrained'	1 sextuplet	1.3	0.6	0.6	0.5	0.5
	1 octuplet	1.7	0.7	0.7	0.6	0.6
	1 decuplet	2.2	0.8	0.8	0.7	0.7
Variant ii: 'typical'	1 sextuplet	0.9	0.5	0.5	0.5	0.5
	1 octuplet	1.1	0.6	0.6	0.6	0.6
	1 decuplet	1.4	0.7	0.7	0.7	0.7
Variant iii: 'deviant'	1 sextuplet	1.6	0.6	0.6	0.5	0.5
	1 octuplet	2.2	0.7	0.7	0.7	0.7
	1 decuplet	2.7	0.8	0.8	0.7	0.7
Variant iv: 'deviant'	1 sextuplet	1.6	0.6	0.6	0.5	0.5
	1 octuplet	2.2	0.7	0.7	0.7	0.7
	1 decuplet	2.8	0.8	0.8	0.7	0.7
		Scenario 2				
Variant i: 'unconstrained'	16 doublets	1.3	1.4	1.3	0.9	0.9
	40 doublets	2.1	2.2	2.1	1.5	1.5
	79 doublets	2.9	3.1	2.9	2.2	2.2
Variant ii: 'typical'	16 doublets	1.3	1.3	1.3	0.9	0.9
	40 doublets	2.7	2.7	2.6	1.5	1.5
	79 doublets	4.8	4.9	4.7	2.2	2.2
Variant iii: 'deviant'	16 doublets	1.8	1.8	3.1	0.9	0.9
	40 doublets	3.3	3.6	7.4	1.5	1.5
	79 doublets	5.4	6.5	14.7	2.1	2.1
Variant iv: 'deviant'	16 doublets	1.6	1.7	2.6	0.9	0.9
	40 doublets	2.8	3.0	6.1	1.5	1.5
	79 doublets	4.2	5.3	11.9	2.1	2.1
		Scenario 3				
Variant i: 'unconstrained'	7 quintuplets	2.7	1.4	2.5	1.2	1.2
	16 quintuplets	4.1	2.2	4.1	2.0	2.0
	31 quintuplets	5.8	3.1	5.8	2.7	2.7
Variant ii: 'typical'	7 quintuplets	2.3	1.4	2.1	1.3	1.3
	16 quintuplets	4.2	2.1	4.0	1.9	1.9
	31 quintuplets	7.3	3.3	7.1	2.7	2.7
Variant iii: 'deviant'	7 quintuplets	3.5	1.5	3.0	1.2	1.2
	16 quintuplets	6.1	2.5	6.3	2.0	2.0
	31 quintuplets	9.0	3.8	11.6	2.8	2.8
Variant iv: 'deviant'	7 quintuplets	3.6	1.5	2.6	1.2	1.2
	16 quintuplets	5.5	2.4	5.3	1.9	1.9
	31 quintuplets	7.9	3.5	9.7	2.7	2.7

Figure 4: Box and whiskers diagrams of Dfbetas in Scenario 2, Variant i, for $N = 500$ and $N = 5,000$.





a – 'Naive' estimation; b – Drop all duplicates; c – Flag and control; d – Drop superfluous duplicates; e – Weighted regression.

Notes: The duplicate records are randomly drawn from the overall distribution.
Box and whiskers show the distribution of Dfbetas (across 2,500 replications) for each of the coefficients in the model (the constant (*_cons*), β_x (x) , β_z (z) and β_t (t)) and for the solutions a to e.

3.4.2 Typical and deviant cases defined on the basis of the distribution of the x variable

To check the robustness of our findings, we follow the same scheme to analyze how the position of the duplicates on the distribution of the independent variable x (rather than the dependent variable y) affects regression estimates. Results are consistent with those presented above, and are available upon request.

4 Conclusions

Reliable data are a prerequisite for well grounded analyses. In this paper we focused on the consequences of duplicate records for regression estimates. A review of the literature shows that there are no papers dealing with this topic. Yet, two recent independent studies by Slomczynski et al. (2015) and by Kuriakose and Robbins (2015) raised the awareness about the quality of survey data and warned about the possible consequences of ignoring the presence of duplicate records. The two teams of researchers showed that a number of broadly used surveys is affected by duplicate records to varying degrees. Unfortunately, little is known about the bias and efficiency loss induced by duplicates in survey data. Present paper fills this gap by addressing two research questions: first, how do duplicates affect regression estimates? Second, how effective are the possible solutions to deal with duplicates?

To this aim we create an artificial data set of $N = 1,500$ observations and four variables with a known covariance matrix. We adopt a Monte Carlo simulation with $2,500$ replications to investigate the consequences of 36 patterns (3 scenarios $\cdot$ 3 cases in each scenario $\cdot$ 4 variants) of duplicate records. The scenarios include: (1) multiple duplications of a single record: sextuplet, octuplet and decuplet; (2) multiple doublets (16, 40, 79, corresponding to 2%, 5%, and 10% of the sample); and (3) multiple quintuplets (7, 16, 31, corresponding to 2%, 5%, and 10% of the sample). The four variants allow us to investigate whether the risk of obtaining biased estimates changes when the duplicates are situated in specific parts of the data distribution: (*i*) on the whole distribution, (*ii*) around the median, (*iii*) on the lower tie, and (*iv*) on the upper tie. For each of the scenarios we run a 'naive' estimation, which ignores the presence of duplicate records.

We investigate the consequences of the duplicates by considering the standardized bias, the risk of obtaining biased estimates, and the efficiency loss to understand when the presence of duplicates is problematic. Finally, we investigate four alternative solutions which can reduce the effect of duplicates on estimation results: (*b*) dropping all duplicates from the sample; (*c*) flagging duplicates and controlling for them in the estimation; (*d*) dropping all superfluous duplicates; (*e*) weighting the observations by the inverse of the duplicates' multiplicity.

Results show that duplicates induce a bias which in our simulations reached the highest value of about 1.75 standard errors. The bias increases with the share of duplicates, and it is lowest when the duplicates are drawn from the overall distribution.

The risk of obtaining biased estimates increases with the share of duplicates in the data: for 7 quintuplets (equivalent to 2.3% of the sample) the risk of obtaining biased estimates is 11%; for 31 quintuplets (10.3% of the sample) the risk of obtaining biased estimates is 44%. The risk of obtaining biased estimates is also higher when they are located on the ties of the distribution. For example, for 31 quintuplets located on the lower or upper tie the probability of obtaining biased estimates is 58.5%. The pattern of duplicates matters: keeping constant the share of duplicates, quintuplets bring a higher probability of biased estimates than doublets. For instance, keeping the share of duplicates constant at 5.3%, 16 quintuplets produce a 28% risk of obtaining biased estimates, whereas 40 doublets carry a risk of 3.5%. Finally, the loss of efficiency due to duplicated records is rather small, reaching the maximum of 20%.

The numerosity and patterns of duplicate records used in this analysis are consistent with those identified by Slomczynski et al. (2015), and they can, therefore, be regarded as realistic. Hence, our first conclusion is that although the bias and efficiency loss related to duplicate records are small, duplicates

create a considerable risk of obtaining biased estimates. Thus, researchers who use data with duplicate records risk to reach misleading conclusions.

Our second conclusion is that weighting the duplicates by the inverse of their multiplicity or dropping the superfluous duplicates are the best solutions among the considered ones. These solutions outperform all the others in reducing the standardized bias, in reducing the risk of obtaining biased estimates, and minimizing the RMSE. The highest risk of biased estimates obtained with these solutions is 8.5% (for 31 quintuplets, i.e. 10.3% of the sample, with duplicates located on the ties), whereas it is 58.5% if the duplicates are neglected. Flagging duplicates and controlling for them is consistently a worst solution, and in some cases (especially for Scenario 2) produces higher bias, higher risk of obtaining biased estimates, and greater efficiency loss than the 'naive' estimation.

Our results do not depend on the sample size we chose ($N = 1,500$): they do not change whether we use a smaller ($N = 500$) or a larger ($N = 5,000$) sample. Similarly, the results do not change if the variants are defined on the basis of one of the independent variables in the regression rather than the dependent one.

These are the first results documenting the effect of duplicates for survey research, and they pave the road for further research on the topic. For instance, our study considers an ideal case when the model used by researchers perfectly fits the relationship in the data, i.e. all relevant predictors are included in the model. This is clearly an unusual situation in social research. Second, our study does not account for heterogeneity of populations. We analyze a case when the relationships of interest are the same for all respondents. In other words, we consider a situation without unmodeled interactions among variables. Third, in our model the records which are substituted by duplicates (the interviews which would have been conducted if no duplicates were introduced in the data) were selected randomly. In reality this is probably not the case, as these are likely the respondents who are most difficult to reach by interviewers. Plausibly, the omitted variables and heterogeneity of population, as well as non-random choice of the interviews replaced by duplicates, may exacerbate the impact of duplicates on regression coefficients. This suggests that our estimates of the effect of duplicates on standardized bias, risk of obtaining biased estimates, and efficiency loss are in many aspects conservative. Moreover, our study assumed that non-unique records were duplicates of true interviews and not purposefully generated fakes. This possibility is also a promising path for future research.

Overall, our results emphasize the importance of collecting 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 duplicates in the data and more refined statistical tools to minimize the consequent estimation bias, risk of obtaining biased estimates, and efficiency loss.

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.

Diekmann, A. (2005). ≪Betrug und Täuschung in der Wissenschaft≫. Datenfälschung, Diagnoseverfahren, Konsequenzen. *Schweizerische Zeitschrift für Soziologie*, 31(1):7–29.

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.

European Social Survey (2015). European Social Survey round 6. Electronic dataset, Bergen: Norwegian Social Science Data Services.

Ferrarini, A. (2011). A fitter use of Monte Carlo simulations in regression models. *Computational Ecology and Software*, 1(4):240–243.

Finn, A. and Ranchhod, V. (2013). Genuine fakes: The prevalence and implications of fieldworker fraud in a large South African survey. Working Paper 115, Southern Africa Labour and Development Research Unit.

Fishman, G. (2005). *A first course in Monte Carlo*. Duxbury Press.

Hassanzadeh, O. and Miller, R. J. (2009). Creating probabilistic databases from duplicated data. *The VLDB Journal–The International Journal on Very Large Data Bases*, 18(5):1141–1166.

Hill, T. P. (1999). The difficulty of faking data. *Chance*, 12(3):27–31.

Koczela, S., Furlong, C., McCarthy, J., and Mushtaq, A. (2015). Curbstoning and beyond: Confronting data fabrication in survey research. *Statistical Journal of the IAOS*, 31(3):413–422.

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.

Schnell, R. (1991). Der Einfluß gefälschter Interviews auf Survey-Ergebnisse. *Zeitschrift für Soziologie*, 20(1):25–35.

Schräpler, J.-P. and Wagner, G. G. (2005). Characteristics and impact of faked interviews in surveys–An analysis of genuine fakes in the raw data of SOEP. *Allgemeines Statistisches Archiv*, 89(1):7–20.

Schreiner, I., Pennie, K., and Newbrough, J. (1988). Interviewer falsification in Census Bureau surveys. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pages 491–496.

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 data sets.

Table 5: Descriptive statistics for the initial data set and for exemplary simulated data sets.

	N. of duplicates	Variables	mean	sd	min	max	obs	missing
Initial data set	0	y	5.213	2.588	−3.878	14.02	1500	0
		x	48.04	16.86	−14.13	99.64	1500	0
		z	4.916	2.402	−3.839	13.99	1500	0
		t	40.03	13.35	−5.743	90.41	1500	0
Scenario 1	6	y	5.212	2.587	−3.878	14.02	1500	0
		x	47.99	16.83	−14.13	99.64	1500	0
		z	4.911	2.400	−3.839	13.99	1500	0
		t	40.01	13.33	−5.743	90.41	1500	0
		duplicates (flag)	0.00400	0.0631	0	1	1500	0
	8	y	5.225	2.588	−3.878	14.02	1500	0
		x	47.96	16.89	−14.13	99.64	1500	0
		z	4.930	2.403	−3.839	13.99	1500	0
		t	39.90	13.43	−5.743	90.41	1500	0
		duplicates (flag)	0.00533	0.0729	0	1	1500	0
	10	y	5.187	2.595	−3.878	14.02	1500	0
		x	47.93	16.87	−14.13	99.64	1500	0
		z	4.909	2.393	−3.839	13.99	1500	0
		t	39.98	13.28	−5.743	90.41	1500	0
		duplicates (flag)	0.00667	0.0814	0	1	1500	0
Scenario 2	16	y	5.217	2.582	−3.878	14.02	1500	0
		x	48.06	16.92	−14.13	99.64	1500	0
		z	4.933	2.409	−3.839	13.99	1500	0
		t	40.00	13.32	−5.743	90.41	1500	0
		duplicates (flag)	0.0213	0.145	0	1	1500	0
	40	y	5.219	2.599	−3.878	14.02	1500	0
		x	48.18	16.81	−14.13	99.64	1500	0
		z	4.929	2.410	−3.839	13.99	1500	0
		t	39.94	13.44	−5.743	90.41	1500	0
		duplicates (flag)	0.0533	0.225	0	1	1500	0
	79	y	5.227	2.582	−3.878	14.02	1500	0
		x	47.99	16.94	−14.13	99.64	1500	0
		z	4.896	2.404	−3.839	13.99	1500	0
		t	40.01	13.42	−5.743	90.41	1500	0
		duplicates (flag)	0.105	0.307	0	1	1500	0
Scenario 3	7	y	5.219	2.584	−3.878	14.02	1500	0
		x	48.09	16.87	−14.13	99.64	1500	0
		z	4.932	2.404	−3.839	13.99	1500	0
		t	39.81	13.47	−5.743	90.41	1500	0
		duplicates (flag)	0.0233	0.151	0	1	1500	0
	16	y	5.240	2.631	−3.878	14.02	1500	0
		x	48.08	16.71	−14.13	99.64	1500	0
		z	4.918	2.453	−3.839	13.99	1500	0
		t	39.91	13.44	−5.743	90.41	1500	0
		duplicates (flag)	0.0533	0.225	0	1	1500	0
	31	y	5.198	2.598	−3.878	14.02	1500	0
		x	48.15	16.61	−14.13	97.58	1500	0
		z	4.941	2.458	−3.839	13.99	1500	0
		t	39.71	13.39	−5.743	90.41	1500	0
		duplicates (flag)	0.103	0.304	0	1	1500	0

B Standardized bias for the remaining coefficients

Table 6: Standardized bias of the β_z coefficient (expressed as a percentage of a standard error)

Variant		Solution:				
		a 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
			Scenario 1			
Variant i: *'unconstrained'*	1 sextuplet	0.5	4.2	4.2	3.9	3.9
	1 octuplet	1.7	−0.8	−0.8	−0.2	−0.2
	1 decuplet	−3.4	−0.0	−0.0	−1.1	−1.1
Variant ii: *'typical'*	1 sextuplet	−20.3	7.5	7.5	1.2	1.2
	1 octuplet	−20.4	4.5	4.5	−0.2	−0.2
	1 decuplet	−24.5	2.2	2.2	−2.6	−2.6
Variant iii: *'deviant'*	1 sextuplet	11.2	−3.9	−3.9	1.8	1.8
	1 octuplet	4.8	−5.3	−5.3	−2.9	−2.9
	1 decuplet	11.1	−1.8	−1.8	2.7	2.7
Variant iv: *'deviant'*	1 sextuplet	12.1	−4.9	−4.9	1.8	1.8
	1 octuplet	10.3	−6.9	−6.9	−2.1	−2.1
	1 decuplet	13.1	−2.7	−2.7	2.9	2.9
			Scenario 2			
Variant i: *'unconstrained'*	16 doublets	0.5	−4.7	0.1	−2.9	−2.9
	40 doublets	−0.2	1.4	−1.0	0.8	0.8
	79 doublets	0.2	0.3	0.8	0.4	0.4
Variant ii: *'typical'*	16 doublets	−50.3	52.5	−44.0	0.3	0.3
	40 doublets	−81.3	79.7	−77.9	−2.5	−2.5
	79 doublets	−124.4	109.7	−121.8	−9.5	−9.5
Variant iii: *'deviant'*	16 doublets	27.7	−29.2	−61.3	−0.2	−0.2
	40 doublets	51.4	−46.8	−104.8	7.1	7.1
	79 doublets	68.0	−75.5	−165.5	6.7	6.7
Variant iv: *'deviant'*	16 doublets	42.2	−41.8	−84.6	0.7	0.7
	40 doublets	70.1	−65.8	−135.1	7.3	7.3
	79 doublets	98.9	−102.2	−206.0	13.0	13.0
			Scenario 3			
Variant i: *'unconstrained'*	7 quintuplets	3.4	3.8	3.2	4.9	4.9
	16 quintuplets	−1.5	−5.0	−1.8	−5.2	−5.2
	31 quintuplets	1.7	2.2	2.3	2.7	2.7
Variant ii: *'typical'*	7 quintuplets	−52.1	19.1	−45.0	1.4	1.4
	16 quintuplets	−81.9	23.8	−77.1	−3.9	−3.9
	31 quintuplets	−120.7	40.7	−117.9	−0.7	−0.7
Variant iii: *'deviant'*	7 quintuplets	22.9	−11.3	−34.3	2.0	2.0
	16 quintuplets	31.8	−21.8	−60.1	−2.1	−2.1
	31 quintuplets	45.2	−27.3	−84.7	2.7	2.7
Variant iv: *'deviant'*	7 quintuplets	35.1	−19.1	−42.8	2.4	2.4
	16 quintuplets	52.9	−32.5	−67.9	1.1	1.1
	31 quintuplets	62.8	−39.8	−103.8	3.1	3.1

Table 7: Standardized bias of the β_t coefficient (expressed as a percentage of a standard error)

Variant		Solution:				
		a 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
		Scenario 1				
Variant i: *'unconstrained'*	1 sextuplet	−2.6	−0.6	−0.6	−1.5	−1.5
	1 octuplet	0.1	−0.3	−0.3	−0.2	−0.2
	1 decuplet	1.4	1.2	1.2	1.6	1.6
Variant ii: *'typical'*	1 sextuplet	−7.8	6.7	6.7	3.7	3.7
	1 octuplet	−11.9	3.9	3.9	1.1	1.1
	1 decuplet	−16.8	4.7	4.7	1.1	1.1
Variant iii: *'deviant'*	1 sextuplet	4.3	1.1	1.1	3.2	3.2
	1 octuplet	4.2	−0.2	−0.2	1.7	1.7
	1 decuplet	9.6	−4.9	−4.9	−0.8	−0.8
Variant iv: *'deviant'*	1 sextuplet	8.3	−0.7	−0.7	3.4	3.4
	1 octuplet	7.3	−1.6	−1.6	1.6	1.6
	1 decuplet	5.4	−2.8	−2.8	−0.5	−0.5
		Scenario 2				
Variant i: *'unconstrained'*	16 doublets	−2.9	1.3	−2.0	−1.3	−1.3
	40 doublets	3.6	0.9	3.5	3.2	3.2
	79 doublets	−1.8	−2.2	−1.6	−2.8	−2.8
Variant ii: *'typical'*	16 doublets	−30.6	31.0	−27.0	−0.1	−0.1
	40 doublets	−51.5	45.0	−49.5	−3.9	−3.9
	79 doublets	−72.8	60.1	−71.1	−6.7	−6.7
Variant iii: *'deviant'*	16 doublets	19.8	−16.3	−48.6	3.2	3.2
	40 doublets	27.5	−30.8	−84.9	0.4	0.4
	79 doublets	40.0	−44.8	−131.6	4.9	4.9
Variant iv: *'deviant'*	16 doublets	21.7	−23.5	−48.7	−0.4	−0.4
	40 doublets	38.4	−44.6	−87.5	−1.1	−1.1
	79 doublets	56.6	−61.3	−125.7	5.5	5.5
		Scenario 3				
Variant i: *'unconstrained'*	7 quintuplets	0.1	−1.0	0.6	−0.8	−0.8
	16 quintuplets	−0.7	−0.4	−0.5	−0.7	−0.7
	31 quintuplets	2.3	0.4	2.0	1.4	1.4
Variant ii: *'typical'*	7 quintuplets	−32.0	8.6	−27.4	−2.0	−2.0
	16 quintuplets	−50.3	16.8	−47.3	−0.7	−0.7
	31 quintuplets	−71.8	19.0	−69.9	−4.5	−4.5
Variant iii: *'deviant'*	7 quintuplets	14.6	−7.4	−30.0	1.5	1.5
	16 quintuplets	19.8	−12.0	−44.6	0.7	0.7
	31 quintuplets	26.1	−17.4	−70.2	1.3	1.3
Variant iv: *'deviant'*	7 quintuplets	18.6	−10.1	−28.1	1.0	1.0
	16 quintuplets	30.3	−17.0	−43.6	2.1	2.1
	31 quintuplets	38.0	−24.4	−63.6	1.8	1.81

Table 8: Standardized bias of the intercept (expressed as a percentage of a standard error)

Variant		Solution:				
		a 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
			Scenario 1			
Variant i:	1 sextuplet	−0.3	0.2	0.2	0.0	0.0
'unconstrained'	1 octuplet	−0.2	−0.6	−0.6	−0.6	−0.6
	1 decuplet	−0.1	−2.0	−2.0	−1.9	−1.9
Variant ii:	1 sextuplet	−5.6	−1.7	−1.7	−2.8	−2.8
'typical'	1 octuplet	−3.7	1.6	1.6	0.6	0.6
	1 decuplet	−0.9	−0.2	−0.2	−0.3	−0.3
Variant iii:	1 sextuplet	−14.0	4.2	4.2	−2.9	−2.9
'deviant'	1 octuplet	−11.8	8.1	8.1	2.7	2.7
	1 decuplet	−17.3	9.4	9.4	2.2	2.2
Variant iv:	1 sextuplet	15.1	−11.0	−11.0	−2.9	−2.9
'deviant'	1 octuplet	17.4	−5.2	−5.2	2.4	2.4
	1 decuplet	15.1	−4.2	−4.2	1.9	1.9
			Scenario 2			
Variant i:	16 doublets	0.2	−0.4	−0.3	−0.1	−0.1
'unconstrained'	40 doublets	−1.2	−0.5	−1.3	−1.2	−1.2
	79 doublets	1.7	3.3	1.9	3.4	3.4
Variant ii:	16 doublets	−4.1	1.5	−1.8	−1.5	−1.5
'typical'	40 doublets	−4.5	6.1	−0.8	1.0	1.0
	79 doublets	−6.0	10.4	−0.9	2.6	2.6
Variant iii:	16 doublets	−49.1	43.9	10.6	−2.0	−2.0
'deviant'	40 doublets	−81.6	72.7	22.1	−6.0	−6.0
	79 doublets	−122.1	108.8	38.4	−10.4	−10.4
Variant iv:	16 doublets	53.4	−51.7	−75.5	1.6	1.6
'deviant'	40 doublets	85.3	−80.1	−125.5	3.6	3.6
	79 doublets	121.8	−118.8	−194.5	5.5	5.5
			Scenario 3			
Variant i:	7 quintuplets	−0.3	−3.4	0.0	−3.1	−3.1
'unconstrained'	16 quintuplets	1.4	0.7	1.6	1.3	1.3
	31 quintuplets	−2.1	−1.4	−2.2	−2.2	−2.2
Variant ii:	7 quintuplets	−6.5	1.7	−3.4	−0.4	−0.4
'typical'	16 quintuplets	−5.9	0.7	−3.3	−1.0	−1.0
	31 quintuplets	−7.1	4.4	−4.7	1.8	1.8
Variant iii:	7 quintuplets	−40.2	20.7	−13.4	−2.3	−2.3
'deviant'	16 quintuplets	−58.6	31.6	−22.8	−1.1	−1.1
	31 quintuplets	−86.0	42.9	−30.3	−4.4	−4.4
Variant iv:	7 quintuplets	39.9	−24.0	−35.4	−0.3	−0.3
'deviant'	16 quintuplets	58.0	−33.2	−59.5	0.4	0.4
	31 quintuplets	93.8	−46.6	−82.5	7.3	7.3

C Root mean square error for the remaining coefficients

Table 9: Normalized RMSE for the β_z coefficient

Variant	Scenario	Solution: *a* 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
		Scenario 1				
Variant i: 'unconstrained'	1 sextuplet	2.1	1.0	1.0	0.9	0.9
	1 octuplet	2.9	1.1	1.1	1.0	1.0
	1 decuplet	3.5	1.3	1.3	1.2	1.2
Variant ii: 'typical'	1 sextuplet	1.3	0.9	0.9	0.9	0.9
	1 octuplet	1.7	1.0	1.0	1.0	1.0
	1 decuplet	2.2	1.2	1.2	1.2	1.2
Variant iii: 'deviant'	1 sextuplet	2.6	1.0	1.0	0.9	0.9
	1 octuplet	3.6	1.1	1.1	1.0	1.0
	1 decuplet	4.4	1.3	1.3	1.1	1.1
Variant iv: 'deviant'	1 sextuplet	2.8	1.0	1.0	0.9	0.9
	1 octuplet	3.6	1.2	1.2	1.0	1.0
	1 decuplet	4.7	1.2	1.2	1.1	1.1
		Scenario 2				
Variant i: 'unconstrained'	16 doublets	2.2	2.2	2.2	1.5	1.5
	40 doublets	3.5	3.6	3.5	2.5	2.5
	79 doublets	4.9	5.2	4.9	3.6	3.6
Variant ii: 'typical'	16 doublets	2.0	2.0	1.9	1.6	1.6
	40 doublets	3.6	3.7	3.6	2.6	2.6
	79 doublets	6.0	6.2	6.0	3.6	3.6
Variant iii: 'deviant'	16 doublets	2.6	2.6	2.9	1.6	1.6
	40 doublets	4.2	4.4	5.5	2.4	2.4
	79 doublets	6.2	6.8	9.5	3.5	3.5
Variant iv: 'deviant'	16 doublets	2.7	2.9	3.2	1.6	1.6
	40 doublets	4.7	5.0	6.3	2.4	2.4
	79 doublets	7.6	8.3	11.4	3.5	3.5
		Scenario 3				
Variant i: 'unconstrained'	7 quintuplets	4.6	2.3	4.3	2.1	2.1
	16 quintuplets	7.0	3.5	6.9	3.1	3.1
	31 quintuplets	9.5	5.1	9.4	4.5	4.5
Variant ii: 'typical'	7 quintuplets	3.2	2.2	3.0	2.1	2.1
	16 quintuplets	5.6	3.5	5.5	3.2	3.2
	31 quintuplets	8.8	5.0	8.6	4.4	4.4
Variant iii: 'deviant'	7 quintuplets	5.6	2.5	3.3	2.1	2.1
	16 quintuplets	8.5	3.9	5.6	3.2	3.2
	31 quintuplets	11.7	5.5	8.5	4.5	4.5
Variant iv: 'deviant'	7 quintuplets	6.0	2.5	3.4	2.1	2.1
	16 quintuplets	9.6	4.0	5.7	3.2	3.2
	31 quintuplets	13.2	5.8	9.2	4.5	4.5

Table 10: Normalized RMSE for the β_t coefficient

Variant	Scenario	Solution:				
		a 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
		Scenario 1				
Variant i: 'unconstrained'	1 sextuplet	3.4	1.6	1.6	1.5	1.5
	1 octuplet	4.9	1.8	1.8	1.7	1.7
	1 decuplet	6.2	2.2	2.2	2.0	2.0
Variant ii: 'typical'	1 sextuplet	2.3	1.6	1.6	1.5	1.5
	1 octuplet	2.9	1.8	1.8	1.8	1.8
	1 decuplet	3.7	2.1	2.1	2.0	2.0
Variant iii: 'deviant'	1 sextuplet	4.6	1.8	1.8	1.5	1.5
	1 octuplet	6.4	2.0	2.0	1.8	1.8
	1 decuplet	8.1	2.2	2.2	2.0	2.0
Variant iv: 'deviant'	1 sextuplet	4.5	1.7	1.7	1.5	1.5
	1 octuplet	5.9	2.0	2.0	1.8	1.8
	1 decuplet	7.9	2.2	2.2	2.0	2.0
		Scenario 2				
Variant i: 'unconstrained'	16 doublets	3.8	3.8	3.8	2.7	2.7
	40 doublets	6.1	6.2	6.1	4.4	4.4
	79 doublets	8.7	8.7	8.7	6.2	6.2
Variant ii: 'typical'	16 doublets	3.2	3.2	3.1	2.7	2.7
	40 doublets	5.3	5.4	5.3	4.3	4.3
	79 doublets	8.1	8.6	8.1	6.2	6.2
Variant iii: 'deviant'	16 doublets	4.4	4.6	4.5	2.6	2.6
	40 doublets	7.1	7.5	8.0	4.3	4.3
	79 doublets	10.0	11.1	13.4	6.0	6.0
Variant iv: 'deviant'	16 doublets	4.4	4.4	4.7	2.6	2.6
	40 doublets	7.1	7.5	8.8	4.3	4.3
	79 doublets	10.2	11.4	14.1	6.0	6.0
		Scenario 3				
Variant i: 'unconstrained'	7 quintuplets	7.9	4.0	7.5	3.7	3.7
	16 quintuplets	11.9	6.2	11.7	5.5	5.5
	31 quintuplets	16.6	8.7	16.3	7.7	7.7
Variant ii: 'typical'	7 quintuplets	5.2	3.8	5.0	3.6	3.6
	16 quintuplets	8.5	5.9	8.3	5.6	5.6
	31 quintuplets	12.3	8.4	12.2	7.8	7.8
Variant iii: 'deviant'	7 quintuplets	9.8	4.2	5.4	3.5	3.5
	16 quintuplets	14.8	6.7	8.7	5.4	5.4
	31 quintuplets	19.9	9.5	13.2	7.8	7.8
Variant iv: 'deviant'	7 quintuplets	9.5	4.2	5.9	3.5	3.5
	16 quintuplets	14.8	6.6	9.2	5.5	5.5
	31 quintuplets	20.4	9.5	14.1	7.6	7.6

Table 11: Normalized RMSE for the intercept

Variant	Scenario	Solution:				
		a 'Naive' estimation	*b* Drop all	*c* Flag and control	*d* Drop superfluous	*e* Weighted regression
		Scenario 1				
Variant i: 'unconstrained'	1 sextuplet	0.8	0.4	0.4	0.4	0.4
	1 octuplet	1.1	0.4	0.4	0.4	0.4
	1 decuplet	1.5	0.5	0.5	0.5	0.5
Variant ii: 'typical'	1 sextuplet	0.5	0.4	0.4	0.4	0.4
	1 octuplet	0.7	0.4	0.4	0.4	0.4
	1 decuplet	0.8	0.5	0.5	0.5	0.5
Variant iii: 'deviant'	1 sextuplet	1.0	0.4	0.4	0.4	0.4
	1 octuplet	1.5	0.5	0.5	0.4	0.4
	1 decuplet	1.8	0.5	0.5	0.5	0.5
Variant iv: 'deviant'	1 sextuplet	1.1	0.4	0.4	0.4	0.4
	1 octuplet	1.4	0.5	0.5	0.4	0.4
	1 decuplet	1.8	0.5	0.5	0.5	0.5
		Scenario 2				
Variant i: 'unconstrained'	16 doublets	0.9	0.9	0.9	0.6	0.6
	40 doublets	1.4	1.4	1.4	1.0	1.0
	79 doublets	2.0	2.0	2.0	1.4	1.4
Variant ii: 'typical'	16 doublets	0.7	0.7	0.7	0.6	0.6
	40 doublets	1.1	1.2	1.1	1.0	1.0
	79 doublets	1.5	1.7	1.5	1.4	1.4
Variant iii: 'deviant'	16 doublets	1.1	1.1	0.9	0.6	0.6
	40 doublets	2.0	2.0	1.4	1.0	1.0
	79 doublets	3.3	3.2	2.0	1.4	1.4
Variant iv: 'deviant'	16 doublets	1.1	1.1	1.2	0.6	0.6
	40 doublets	2.0	2.1	2.5	1.0	1.0
	79 doublets	3.3	3.6	4.6	1.4	1.4
		Scenario 3				
Variant i: 'unconstrained'	7 quintuplets	1.8	0.9	1.7	0.8	0.8
	16 quintuplets	2.8	1.4	2.7	1.3	1.3
	31 quintuplets	3.8	2.0	3.7	1.8	1.8
Variant ii: 'typical'	7 quintuplets	1.2	0.9	1.1	0.8	0.8
	16 quintuplets	1.7	1.3	1.7	1.3	1.3
	31 quintuplets	2.3	1.9	2.2	1.8	1.8
Variant iii: 'deviant'	7 quintuplets	2.4	1.0	1.2	0.8	0.8
	16 quintuplets	3.8	1.6	1.9	1.3	1.3
	31 quintuplets	5.9	2.3	2.7	1.8	1.8
Variant iv: 'deviant'	7 quintuplets	2.4	1.0	1.4	0.8	0.8
	16 quintuplets	3.7	1.6	2.3	1.3	1.3
	31 quintuplets	5.9	2.3	3.5	1.8	1.8