Survey Research Methods (2017)
Vol. 11, No. 1, pp. 17-44
doi:10.18148/srm/2017.v11i1.7149

ISSN 1864-3361
http://www.surveymethods.org

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

Francesco Sarracino
Institut National de la Statistique et des Études Économiques du Grand-Duché du Luxembourg (STATEC)
Luxembourg, Luxembourg
and
LCSR National Research University
Higher School of Economics
Moscow, Russian Federation

Małgorzata Mikucka
Mannheimer Zentrum für Europäische Sozialforschung (MZES)
Mannheim, Germany
and
LCSR National Research University
Higher School of Economics
Moscow, Russian Federation

Recent studies documented that survey data contain duplicate records. In this paper, we assess how duplicate records affect regression estimates, and we evaluate the effectiveness of solutions to deal with them. Results show that duplicates bias the estimated coefficients and standard errors. 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 in reducing the bias and the risk of obtaining biased estimates. However, both solutions overestimate standard errors, reducing the statistical power of estimates. Our study illustrates the risk of using data in presence of duplicate records and call for further research on strategies to analyse affected data.

Keywords: Duplicate records, Estimation Bias, Monte Carlo Simulation, Inference, Survey Data Quality

1 Introduction

To achieve reliable results survey data must accurately report respondents' answers. Yet, sometimes they don't. The study of Slomczynski, Powałko, and Krauze (2017) published in this issue of SRM investigated survey projects widely used in social sciences, and reported a considerable number of duplicate records in 17 out of 22 international 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 originate from the answers of two real respondents. It is more probable that 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, 2004; Diekmann, 2005; Koczela, Furlong, McCarthy, & Mushtaq, 2015; Kuriakose & Robbins, 2016; Waller, 2013).

Contact information: Francesco Sarracino, STATEC, 13 rue Erasme, L-2013 Luxembourg (email: f.sarracino@gmail.com)

1.1 Duplicate records in social survey data

Slomczynski et al. (2017) 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. (2017) 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 (2016) analyzed near duplicates in data sets commonly used in social sciences and showed that 16% of analysed surveys reported a high risk of widespread falsification with near duplicates. The authors emphasized that demographic and geographical variables are rarely falsified, because they usually have to meet the sampling frame. Behavioral and attitudinal variables, on the other hand, were falsified more often. In such cases, interviewers may only copy 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 & 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 & 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 & Robbins, 2016). In turn, this may increase the statistical significance of the coefficients and affect the substantive conclusions.

The implications of duplicates for regression estimates may differ according to the characteristics of the observations being duplicated. Slomczynski et al. (2017) 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, Ipeirotis, & Verykios, 2007; Hassanzadeh & Miller, 2009; Kuriakose & Robbins, 2016; Schreiner, Pennie, & Newbrough, 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 of variables. 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

Our study is the first analysis of 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).

We chose the number of duplicates to mimic the results provided by Slomczynski et al. (2017). 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, consistently with the suggestion by Slomczynski et al. (2017), that Variants *iii* and *iv* affect regression estimates more than Variant *i*, and that Variant *ii* affects them the least. Additionally, we repeat the whole analysis to test the robustness of our findings by checking 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. percentage bias of coefficients;
Measure B. bias of the standard errors;
Measure C. risk of obtaining biased estimates, as measured by Dfbetas;
Measure D. 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. analysing 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.

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

variables	x	z	t
x	1.00		
z	−0.04	1.00	
t	0.09	−0.06	1.00

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 relationships among variables are 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 (derived from data without duplicates) 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*). In the same step we also estimate regression models using the four alternative solutions (b–e) to deal with duplicate records. Finally, we compute the bias of coefficients and standard errors, the risk of obtaining biased estimates, and the Root Mean Square Error to assess the effect of duplicates on regression estimates and the effectiveness of the solutions. 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).

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

$$y_i = 5.36 - 0.04 \cdot x_i + 0.16 \cdot z_i + 0.023 \cdot t_i + \epsilon_i \qquad (1)$$

where the coefficients are also retrieved from the sixth wave of the European Social Survey. All variables and the error term ϵ_i are normally distributed. The descriptive statistics

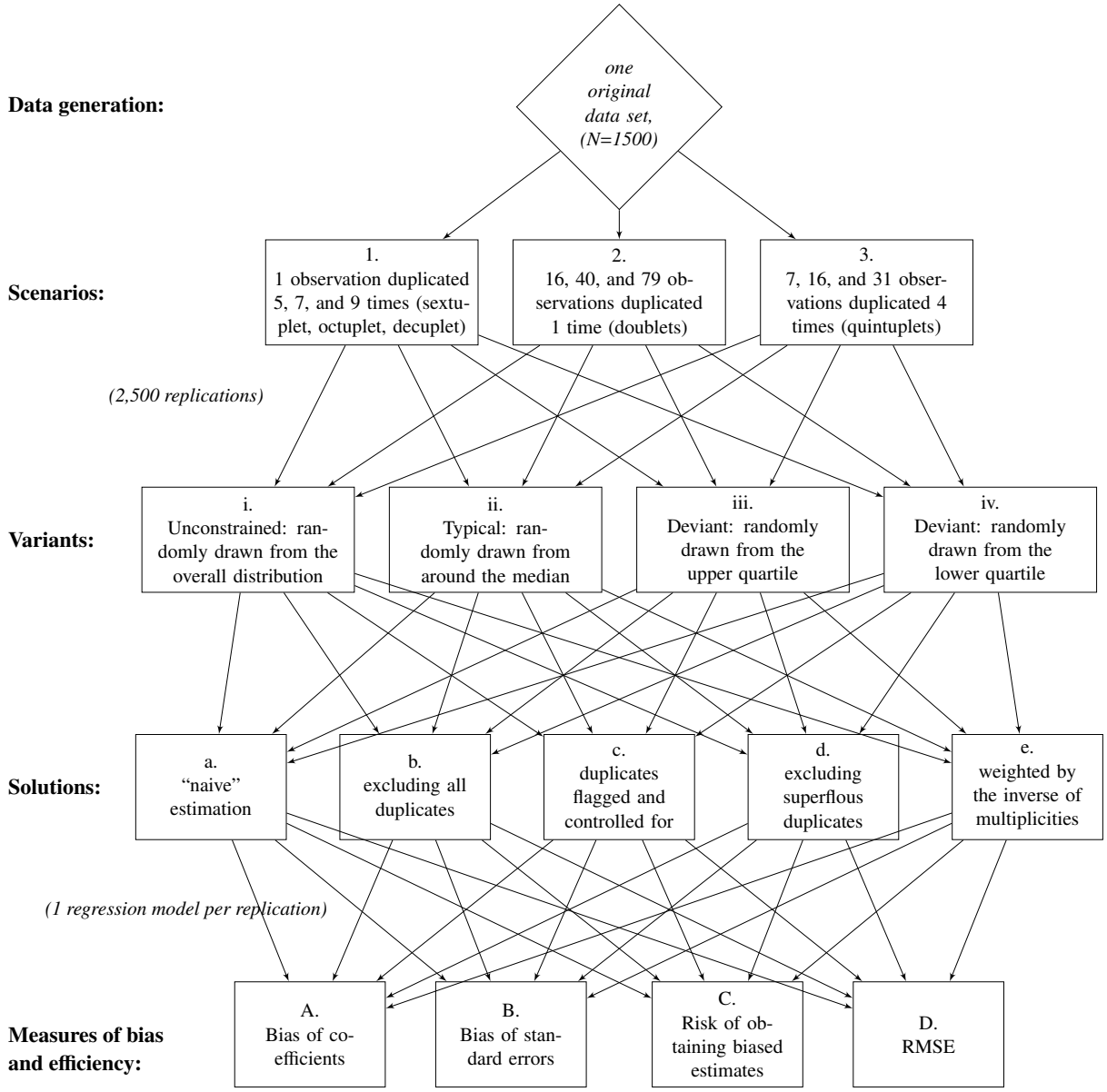

Figure 1. Diagram summarizing the empirical strategy.

of the generated data set are shown in the first four lines of Table A1 in Appendix A.

2.2 Duplicating selected observations

In the second step we use a Monte Carlo simulation to generate duplicate records, which replace for 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 randomly chosen original observations. These cases are possible in the light of the analysis by Slomczynski et al. (2017) who identified in real survey data

instances of octuplets. In this scenario the share of duplicates in the sample is small, ranging from 0,4% for a sextuplet to 0,7% 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 2,1% (16 doublets), 5,3% (40 doublets), and 10,5% (79 doublets). These shares are consistent with the results by Slomczynski et al. (2017), 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 randomly chosen original records respectively. In this scenario the share of duplicate records is 2,3% (7 quintuplets), 5,3% (16 quintuplets), and 10,3% (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 A1 in Appendix A reports the descriptive statistics of some of the data sets produced during the replications (lines 5 to 45).

2.3 "Naive" estimation and alternative solutions

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

$$y_i = \alpha + \beta_x \cdot x_i + \beta_z \cdot z_i + \beta_t \cdot t_i + \varepsilon_i \qquad (2)$$

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 percentage bias, the standard errors, the risk of obtaining biased estimates, and the Root Mean Square Error resulting from the mere presence of duplicate records (see Section 2.4).

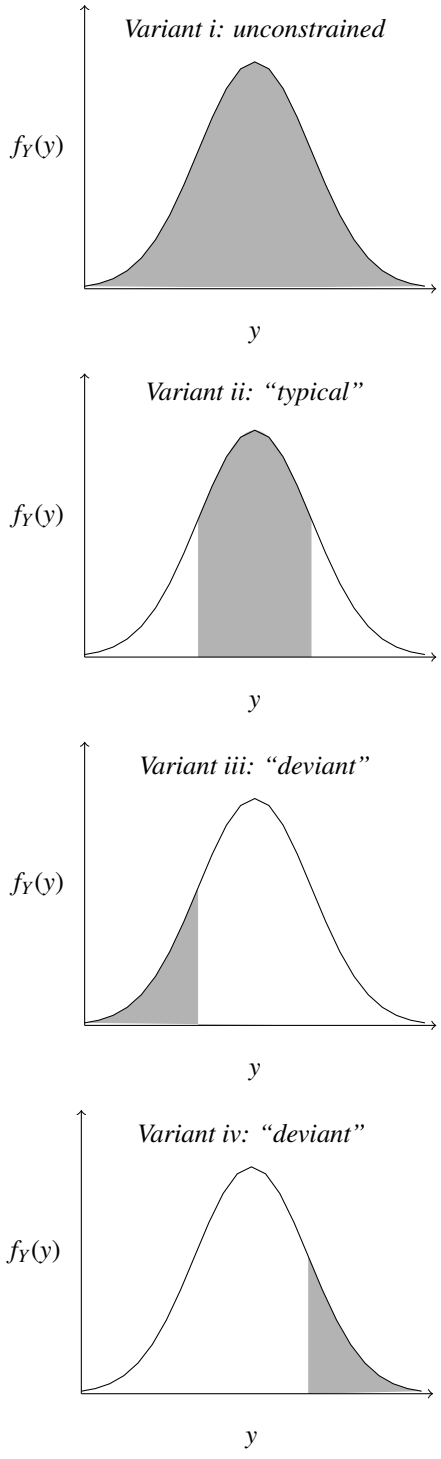

Figure 2. Presentation of the Variants *i–iv* used in the Monte Carlo simulation.

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. Slomczynski et al. (2017) proposed this solution 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 & Robbins, 2016, 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.

Solution e: Weight by the inverse of multiplicities. Lessler and Kalsbeek (1992) proposed this method. 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 four measures to assess the consequences of duplicates for regression estimates, and to investigate the efficiency of the solutions to deal with them.

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

$$\text{Bias of coefficients} = \left(\frac{\overline{\widehat{\beta^i}} - \beta}{\beta} \right) \cdot 100\% \tag{3}$$

where i indicates a specific replication, β is the true coefficient from Equation 1, and $\overline{\widehat{\beta^i}}$ is the average of estimated coefficients ($\widehat{\beta^i}$).

Measure B: Bias of standard errors. To test whether duplicates artificially increase the power of regression estimates, we compute the average of the standard errors ($\overline{\text{SE}(\widehat{\beta^i})}$) for each scenario, variant, and solution. For ease of interpretation, we express our measure as a percentage of the standard errors estimated in the true model (see Equation 4).

$$\text{Bias of S.E.} = \left(\frac{\overline{\text{SE}(\widehat{\beta^i})}}{\text{SE}(\beta)} \right) \cdot 100\% \tag{4}$$

Measure C: Risk of obtaining biased estimates. It is possible to obtain biased estimates even if the average bias is zero. This can happen if the upward and downward biases offset each other. 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 estimated coefficient (see Equation 5).

$$\text{Dfbeta}^i = \frac{\widehat{\beta^i} - \beta}{\text{SE}(\widehat{\beta^i})} \tag{5}$$

Dfbetas measure the bias of a specific estimation, thus, they complement percentage bias by informing about the risk of obtaining biased estimates. The risk is computed according to Equation 6. 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]

$$x^i = \begin{cases} 1 & \text{if } |\text{Dfbeta}^i| > 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

$$Pr(\text{Bias}) = \overline{x^i} \cdot 100\%$$

Measure D: Root Mean Square Error (RMSE). Root mean square error is an overall measure of the quality of the prediction and it reflects both its bias and its dispersion (see Equation 7).

$$\text{RMSE} = \sqrt{\left(\overline{\widehat{\beta^i}} - \beta\right)^2 + \left(\text{SE}(\widehat{\beta^i})\right)^2} \tag{7}$$

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 Percentage bias

Table 2 shows the percentage bias of the coefficient β_x for the considered scenarios, variants, and solutions. The results for the other coefficients are reported in Tables B1–B3 in Appendix B. The third column of Table 2 contains information about the percentage bias for solution *a*, i.e. the "naive" estimation.

Overall, the percentage bias takes values between nearly zero (in Scenario 1 and in all Scenarios in Variant *i*) and about 7%. For β_x it reaches the maximum values of 4% for 79 doublets and 6% for 31 quintuplets. The maximum bias for β_z and β_t is 5%–7%, and for the intercept it is 2.5%–4% (see Appendix B).

The results show some regularities. First, number and composition of duplicates matter. The bias systematically increases with the share of duplicates in the data, and duplicates consisting of quintuplets produce greater bias than duplicates consisting of doublets. Second, the choice of records to be duplicated plays a role. The "unconstrained" variant (Variant *i*), where duplicate cases are randomly selected, produces virtually no bias, even when the share of duplicates reaches 10% of the sample. On the other hand, Variant *ii* produces similar bias as Variants *iii* and *iv*. In other words, contrary to our expectations, the duplication of "typical" records produces a bias similar to the one induced by the presence of "deviant" duplicates. Only randomly chosen duplicates generate no bias. Third, although previous studies suggested that duplicates may introduce "random noise" to the data, thus leading to attenuation bias, we did not find the evidence to support this expectation: depending on the Variant (*ii*–*iv*), for each variable the presence of duplicates induces a mix of overestimated and underestimated coefficients.

Among the four solutions to deal with duplicates, solutions *d* and *e*, i.e. dropping the superfluous duplicates and weighting by the inverse of multiplicity perform the best in all Variants, reducing the bias to zero. On the other hand, dropping all duplicates (solution *b*) and flagging duplicates and controlling for them in the regression (solution *c*) perform poorly. Especially in Scenario 2 both these solutions increase the bias of all coefficients; in Scenario 3 they reduce the bias, but to a lesser degree than solutions *d* and *e*.

In sum, duplicates can systematically bias regression estimates if they are not randomly created. However, the bias in our simulation did not exceed 10% of the true coefficients. Moreover, dropping superfluous duplicates or weighting by the inverse of their multiplicity are effective ways to reduce the bias to zero.

3.2 Standard errors

To understand whether duplicates artificially increase the statistical power of regression estimates, we inspect the average estimated standard errors, as shown in Table 3 for β_x. The results for other coefficients are presented in Tables C1–C3 in Appendix C.

The results show, similarly to the case of percentage bias, that duplicates in Variant *i*, i.e. randomly drawn from the overall distribution ("unconstrained"), do not affect the estimates of the standard errors. In Variant *ii*, in which the duplicates are located around the median of the dependent variable, the estimated standard errors are biased downwards by maximum 2%–3%, thus the confidence intervals are narrower than in the true model. On the contrary, in Variants *iii* and *iv*, i.e. the two "deviant" cases, duplicates lead to standard errors biased upwards by maximum 2%–3%, and to broader confidence intervals. Both effects are stronger when data contain more duplicates, i.e. when data contain 79 doublets or 31 quintuplets.

Among the considered solutions, flagging and controlling for duplicates (Solution *c*) leads to systematically narrower confidence intervals. This is especially worrisome because the same solution produces the most biased coefficients. In other words, this solution may result in biased and significant coefficients, thus affecting the interpretation of the results. The remaining three Solutions, *b*, *d*, and *e*, produce slightly greater standard errors than the naive estimation. The relative performance of the solutions varies across specific coefficients: for β_x and β_z Solution *e* works better than dropping duplicates, but it overestimates the standard errors more than dropping the duplicates for β_t and the intercept.

Summing up, we find no evidence that the duplicates artificially increase the statistical power of the estimates if the duplicates are created randomly. However, if duplicates are chosen from the center of the distribution they may lead to narrower confidence intervals, thus artificially increasing the statistical power. On the other hand, duplication of "deviant" cases reduces the power of estimates. In both cases the effect is small, up to 3% of the true standard errors. Yet, the most effective solutions to reduce the bias of coefficients, i.e. dropping the superfluous duplicates and weighting by the inverse of multiplicity, increase the estimated standard errors, thus reducing the power of estimates.

3.3 Risk of obtaining biased estimates

While percentage bias informs about the average bias due to duplicates, it is plausible that estimates in specific replications have upward and downward biases which, on average, offset each other. In other words, even with moderate bias, researchers can 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

Table 2
Percentage bias of the β_x coefficient (as a percentage of β_x).

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	−0.0	0.0	0.0	0.0	0.0
1 octuplet	0.0	−0.0	−0.0	−0.0	−0.0
1 decuplet	0.0	−0.0	−0.0	−0.0	−0.0
Variant ii: "typical"					
1 sextuplet	−0.3	0.1	0.1	0.0	0.0
1 octuplet	−0.3	0.1	0.1	0.0	0.0
1 decuplet	−0.5	0.1	0.1	0.0	0.0
Variant iii: "deviant"					
1 sextuplet	0.3	−0.1	−0.1	−0.0	−0.0
1 octuplet	0.4	−0.0	−0.0	0.0	0.0
1 decuplet	0.6	−0.0	−0.0	0.0	0.0
Variant iv: "deviant"					
1 sextuplet	0.2	−0.0	−0.0	−0.0	−0.0
1 octuplet	0.3	−0.0	−0.0	0.0	0.0
1 decuplet	0.3	0.0	0.0	0.0	0.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	−0.0	−0.0	−0.0	−0.0	−0.0
40 doublets	0.0	0.0	0.0	0.0	0.0
79 doublets	0.0	0.1	0.0	0.0	0.0
Variant ii: "typical"					
16 doublets	−0.8	0.8	−0.7	−0.0	−0.0
40 doublets	−2.1	2.1	−2.0	−0.1	−0.1
79 doublets	−4.2	4.2	−4.1	−0.2	−0.2
Variant iii: "deviant"					
16 doublets	1.0	−1.0	−2.8	0.0	0.0
40 doublets	2.3	−2.6	−7.1	0.0	0.0
79 doublets	4.5	−5.6	−14.4	0.3	0.3
Variant iv: "deviant"					
16 doublets	0.6	−0.7	−2.1	0.0	0.0
40 doublets	1.5	−1.8	−5.6	0.0	0.0
79 doublets	2.9	−4.0	−11.4	0.1	0.1
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	0.1	−0.1	0.1	−0.0	−0.0
16 quintuplets	0.0	−0.1	0.0	−0.0	−0.0
31 quintuplets	−0.0	−0.0	0.0	−0.0	−0.0
Variant ii: "typical"					
7 quintuplets	−1.4	0.4	−1.2	−0.0	−0.0
16 quintuplets	−3.3	0.8	−3.1	−0.1	−0.1
31 quintuplets	−6.3	1.6	−6.1	−0.1	−0.1
Variant iii: "deviant"					
7 quintuplets	1.5	−0.4	−2.2	0.0	0.0
16 quintuplets	3.4	−1.0	−5.4	0.0	0.0
31 quintuplets	6.2	−2.0	−10.8	0.2	0.2
Variant iv: "deviant"					
7 quintuplets	1.1	−0.3	−1.7	0.0	0.0
16 quintuplets	2.3	−0.7	−4.4	0.0	0.0
31 quintuplets	4.3	−1.3	−8.8	0.2	0.2

Table 3
Average standard error of β_x coefficient (expressed as a percentage of the true standard error of β_x).

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	100.0	100.2	100.0	100.2	99.9
1 octuplet	100.0	100.3	100.0	100.2	100.0
1 decuplet	100.0	100.3	100.0	100.3	100.0
Variant ii: "typical"					
1 sextuplet	99.9	100.2	100.1	100.2	99.9
1 octuplet	99.8	100.3	100.1	100.2	100.0
1 decuplet	99.8	100.4	100.1	100.3	100.1
Variant iii: "deviant"					
1 sextuplet	100.1	100.2	100.0	100.2	99.9
1 octuplet	100.2	100.2	100.0	100.2	100.0
1 decuplet	100.2	100.3	100.0	100.3	100.0
Variant iv: "deviant"					
1 sextuplet	100.1	100.2	100.0	100.2	99.9
1 octuplet	100.2	100.2	100.0	100.2	100.0
1 decuplet	100.2	100.3	100.0	100.3	100.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	100.0	101.1	100.0	100.5	100.0
40 doublets	100.0	102.8	100.0	101.4	100.5
79 doublets	100.0	105.7	100.0	102.7	101.1
Variant ii: "typical"					
16 doublets	99.6	101.5	99.7	100.5	100.2
40 doublets	99.0	103.8	99.1	101.3	100.9
79 doublets	98.0	107.8	98.1	102.6	102.0
Variant iii: "deviant"					
16 doublets	100.4	100.7	99.1	100.5	99.9
40 doublets	100.9	101.8	97.6	101.4	100.0
79 doublets	101.7	103.5	95.1	102.9	100.3
Variant iv: "deviant"					
16 doublets	100.4	100.7	99.0	100.6	99.9
40 doublets	100.9	101.7	97.3	101.4	99.9
79 doublets	101.8	103.3	94.4	102.9	100.1
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	100.0	101.2	100.0	100.9	100.6
16 quintuplets	100.0	102.8	100.0	102.2	101.5
31 quintuplets	100.0	105.6	100.0	104.4	103.2
Variant ii: "typical"					
7 quintuplets	99.3	101.4	99.5	100.9	100.6
16 quintuplets	98.5	103.2	98.6	102.2	101.8
31 quintuplets	97.0	106.4	97.2	104.4	103.8
Variant iii: "deviant"					
7 quintuplets	100.6	101.0	99.3	101.0	100.4
16 quintuplets	101.3	102.4	98.3	102.2	101.2
31 quintuplets	102.4	104.8	96.5	104.5	102.8
Variant iv: "deviant"					
7 quintuplets	100.7	101.0	99.2	101.0	100.4
16 quintuplets	101.4	102.4	98.0	102.2	101.2
31 quintuplets	102.5	104.8	96.0	104.5	102.6

results for Scenario 1 because in this case the risk of obtaining biased estimates is virtually zero. Results, however, are detailed in Table 4. 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, and it 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 4. 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: it is 0.2%–0.7% (depending on the variant) for 16 doublets, but it grows to 14.0%–32.5% when 79 doublets are included in the data. In Scenario 3 it grows from 3.0%–20.2% for 7 quintuplets, to 43.0%–58.6% 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 percentage bias, weighting by the inverse of the multiplicity (Solution *e*), and dropping the superfluous duplicates (Solution *d*) perform better than other solutions (see Table 4 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 dropping all duplicate records decreases the probability of obtaining unbiased estimates.

3.4 Root Mean Square Error

Table 5 shows the values of normalized RMSE for coefficient β_x. Results for other coefficients are available in Tables D1–D3 in Appendix D. Scores are overall small, reaching about 9% of the β_x coefficient, 15% of β_z, 24% of β_t, and 6% of the intercept.

The RMSE captures both the bias and the standard errors, thus it is not surprising that the results are consistent with those presented in the sections above. First, RMSE increases with the number of duplicates, and it is the highest for 79 doublets and 31 quintuplets. Second, the presence of randomly duplicated observations (Variant *i*) has little effect on the efficiency of the estimates, whereas the presence of "typical" (Variant *ii*) and "deviant" (Variant *iii* and *iv*) duplicates reduces the efficiency of estimates.

Consistently with previous results, Solutions *d* and *e*, i.e. dropping superfluous duplicates and weighting the data perform the best, reasonably reducing the RMSE values. In contrast to that, flagging the duplicates and controlling for them (Solution *c*) performs poorly, and in some cases (especially in Scenario 2, but for β_x also in Scenario 3) it further reduces the efficiency of the estimates.

3.5 Robustness

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 (DFbetas) for Scenario 2, Variant *i*. The complete set of results is available upon request.

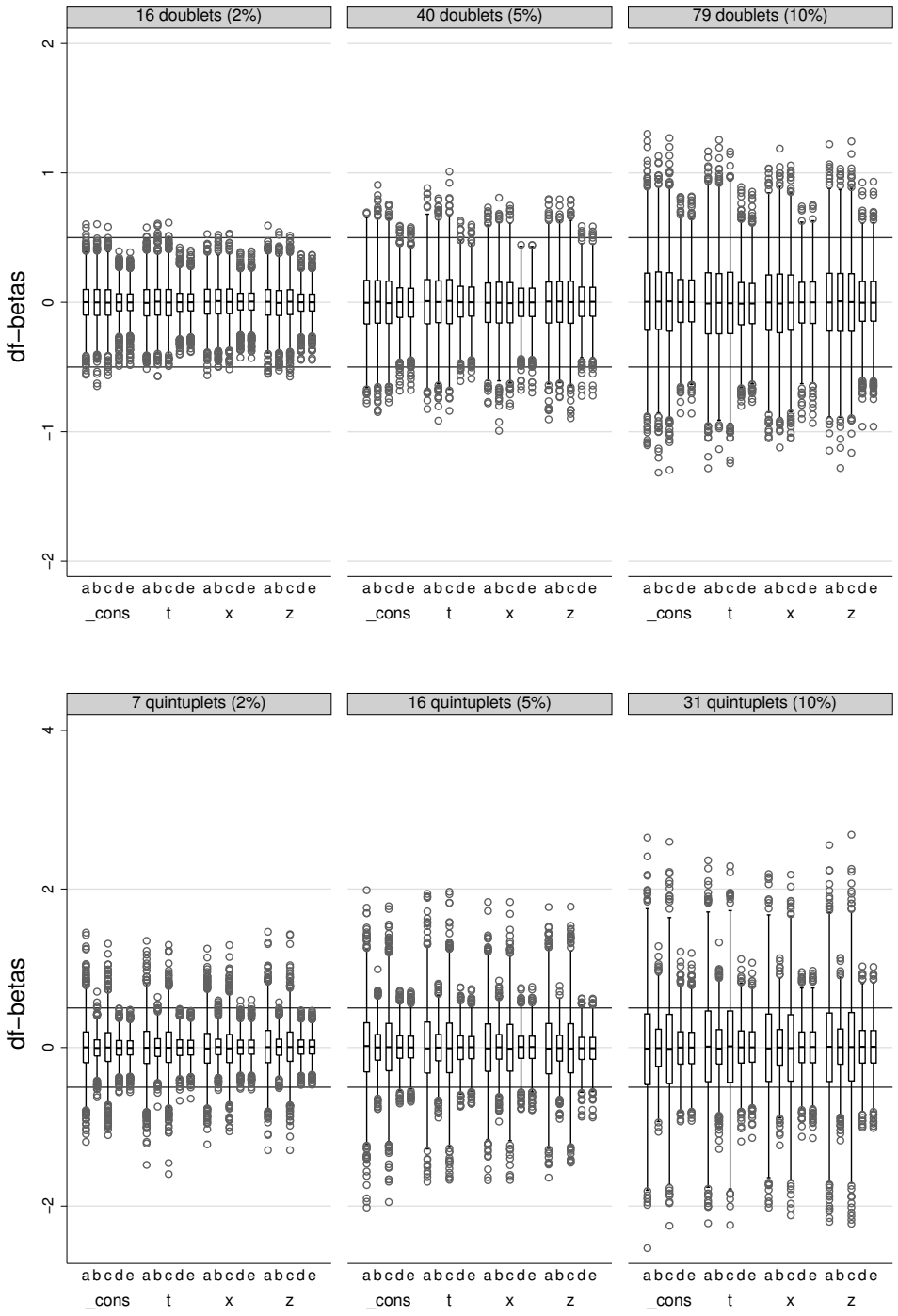

Figure 3. Box and whiskers diagrams of Dfbetas in Scenario 2 and 3, Variant *i*. The duplicate records are randomly drawn from the overall distribution. Box and whiskers show the distribution of Dfbetas (across 2,500 replications) for each of the coefficients in the model and for the solutions a to e.

Notes: a: "Naive" estimation; b: Drop all duplicates; c: Flag and control; d: Drop superfluous duplicates; e: Weighted regression. _cons: regression constant; x: β_x; z: β_z; t: β_t.

Table 4
Probability of obtaining unbiased estimates ($\text{Dfbeta}_i < 0.5$*).*

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	99.2	100.0	100.0	100.0	100.0
1 octuplet	97.0	100.0	100.0	100.0	100.0
1 decuplet	94.4	100.0	100.0	100.0	100.0
Variant ii: "typical"					
1 sextuplet	99.9	100.0	100.0	100.0	100.0
1 octuplet	99.7	100.0	100.0	100.0	100.0
1 decuplet	98.9	100.0	100.0	100.0	100.0
Variant iii: "deviant"					
1 sextuplet	97.9	100.0	100.0	100.0	100.0
1 octuplet	94.7	100.0	100.0	100.0	100.0
1 decuplet	90.9	100.0	100.0	100.0	100.0
Variant iv: "deviant"					
1 sextuplet	98.1	100.0	100.0	100.0	100.0
1 octuplet	94.5	100.0	100.0	100.0	100.0
1 decuplet	89.2	100.0	100.0	100.0	100.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	99.8	99.8	99.8	100.0	100.0
40 doublets	96.5	96.4	96.5	99.6	99.7
79 doublets	86.0	86.9	86.1	96.1	96.4
Variant ii: "typical"					
16 doublets	100.0	100.0	100.0	100.0	100.0
40 doublets	96.3	96.8	96.6	99.6	99.7
79 doublets	77.1	80.5	78.0	96.1	96.5
Variant iii: "deviant"					
16 doublets	99.3	99.2	96.6	100.0	100.0
40 doublets	88.5	86.8	68.3	99.6	99.7
79 doublets	66.5	60.7	44.1	96.8	96.9
Variant iv: "deviant"					
16 doublets	99.3	99.0	97.4	100.0	100.0
40 doublets	89.2	87.1	63.7	99.7	99.7
79 doublets	67.7	61.2	25.9	96.8	96.9
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	89.2	99.7	91.0	99.9	99.9
16 quintuplets	71.9	96.4	72.4	98.0	98.2
31 quintuplets	55.7	87.3	56.4	91.2	91.8
Variant ii: "typical"					
7 quintuplets	97.0	99.8	98.0	99.9	99.9
16 quintuplets	80.7	96.8	82.3	97.7	98.1
31 quintuplets	57.1	87.8	58.5	90.6	91.4
Variant iii: "deviant"					
7 quintuplets	80.3	99.6	95.3	100.0	100.0
16 quintuplets	58.2	93.9	72.2	97.8	98.1
31 quintuplets	41.4	81.5	48.5	90.7	91.0
Variant iv: "deviant"					
7 quintuplets	79.9	99.7	95.2	100.0	100.0
16 quintuplets	57.1	94.1	72.4	97.9	98.1
31 quintuplets	41.9	81.7	44.2	91.4	91.7

Table 5
Normalized RMSE *of the β_x coefficient (in percentage).*

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	9.1	9.1	9.1	9.1	9.1
1 octuplet	9.1	9.1	9.1	9.1	9.1
1 decuplet	9.1	9.1	9.1	9.1	9.1
Variant ii: "typical"					
1 sextuplet	9.1	9.1	9.1	9.1	9.1
1 octuplet	9.1	9.1	9.1	9.1	9.1
1 decuplet	9.1	9.1	9.1	9.1	9.1
Variant iii: "deviant"					
1 sextuplet	9.1	9.1	9.1	9.1	9.1
1 octuplet	9.1	9.1	9.1	9.1	9.1
1 decuplet	9.1	9.1	9.1	9.1	9.1
Variant iv: "deviant"					
1 sextuplet	9.1	9.1	9.1	9.1	9.1
1 octuplet	9.1	9.1	9.1	9.1	9.1
1 decuplet	9.1	9.1	9.1	9.1	9.1
Scenario 2					
Variant i: "unconstrained"					
16 doublets	9.1	9.2	9.1	9.1	9.1
40 doublets	9.1	9.4	9.1	9.2	9.1
79 doublets	9.1	9.6	9.1	9.3	9.2
Variant ii: "typical"					
16 doublets	9.1	9.3	9.1	9.1	9.1
40 doublets	9.2	9.7	9.2	9.2	9.2
79 doublets	9.8	10.7	9.8	9.3	9.3
Variant iii: "deviant"					
16 doublets	9.2	9.2	9.4	9.1	9.1
40 doublets	9.5	9.6	11.4	9.2	9.1
79 doublets	10.3	10.9	16.8	9.4	9.1
Variant iv: "deviant"					
16 doublets	9.2	9.2	9.3	9.1	9.1
40 doublets	9.3	9.4	10.5	9.2	9.1
79 doublets	9.7	10.2	14.3	9.4	9.1
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	9.1	9.2	9.1	9.2	9.1
16 quintuplets	9.1	9.3	9.1	9.3	9.2
31 quintuplets	9.1	9.6	9.1	9.5	9.4
Variant ii: "typical"					
7 quintuplets	9.2	9.2	9.1	9.2	9.2
16 quintuplets	9.6	9.4	9.5	9.3	9.3
31 quintuplets	10.9	9.8	10.8	9.5	9.4
Variant iii: "deviant"					
7 quintuplets	9.3	9.2	9.3	9.2	9.1
16 quintuplets	9.8	9.4	10.5	9.3	9.2
31 quintuplets	11.2	9.7	13.9	9.5	9.4
Variant iv: "deviant"					
7 quintuplets	9.2	9.2	9.2	9.2	9.1
16 quintuplets	9.5	9.3	9.9	9.3	9.2
31 quintuplets	10.3	9.6	12.4	9.5	9.3

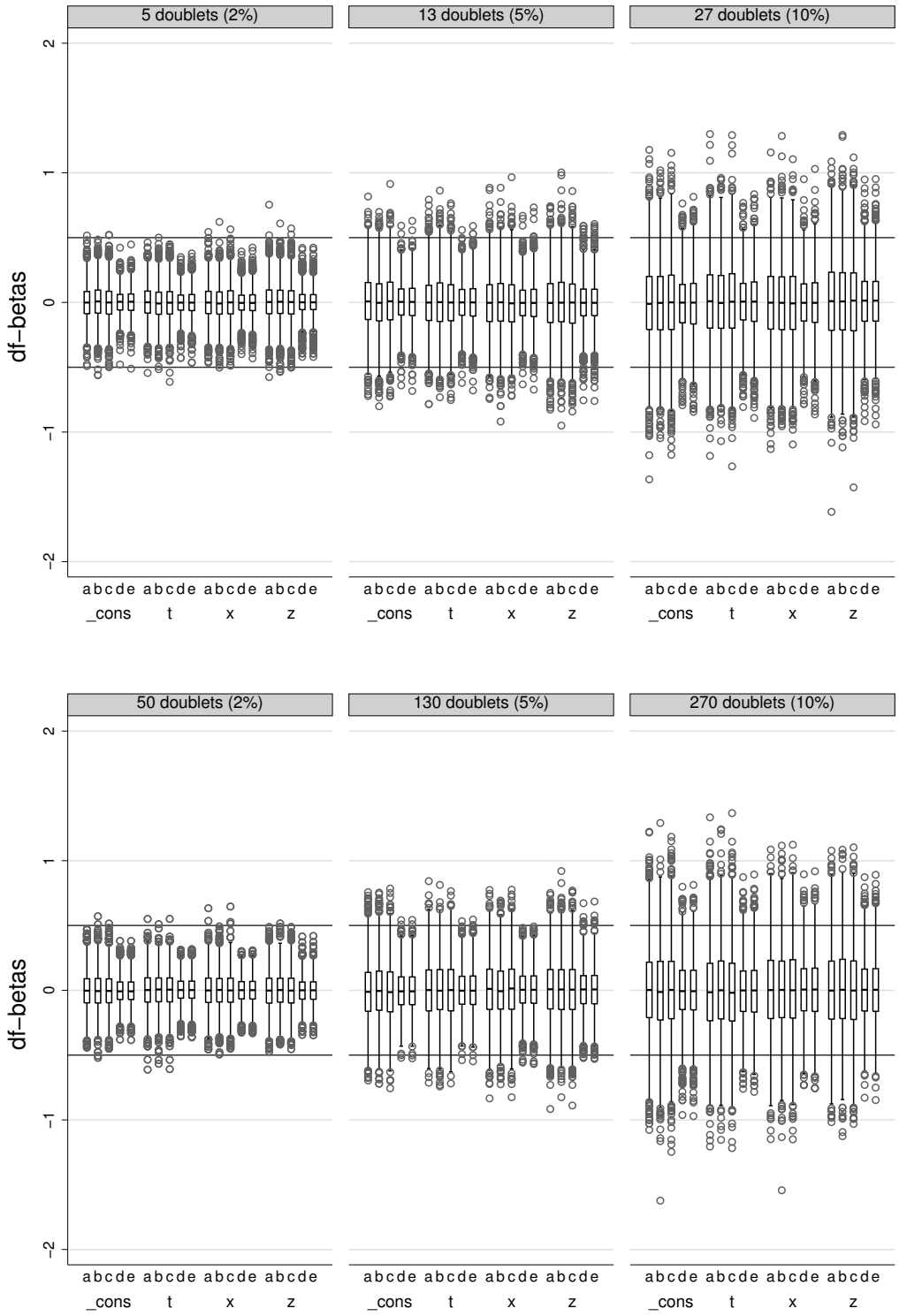

Figure 4. Box and whiskers diagrams of Dfbetas in Scenario 2, Variant *i*, for $N = 500$ and $N = 5,000$. The duplicate records are randomly drawn from the overall distribution. Box and whiskers show the distribution of Dfbetas (across 2,500 replications) for each of the coefficients in the model and for the solutions a to e.

Notes: a: "Naive" estimation; b: Drop all duplicates; c: Flag and control; d: Drop superfluous duplicates; e: Weighted regression. _cons: regression constant; x: β_x; z: β_z; t: β_t.

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.

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. (2017) and by Kuriakose and Robbins (2016) raised the awareness about the quality of survey data and they warned about the possible consequences of ignoring the presence of duplicate records. The two teams of researchers showed that a number of widely 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 partly 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 created an artificial data set of $N = 1,500$ observations and four variables with a known covariance matrix. We adopted a Monte Carlo simulation with 2,500 replications to investigate the consequences of 36 patterns (3 scenarios · 3 cases in each scenario · 4 variants) of duplicate records. The scenarios included: (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 allowed us to investigate whether the reliability of regression estimates changed when the duplicates were 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 of the distribution of the dependent variable.

For each of the scenarios we run a "naive" estimation, which ignored the presence of duplicate records. This allowed us to investigate the consequences of duplicate records for regression estimates. Specifically, we investigated the percentage bias, the standard errors, the risk of obtaining biased estimates, and the root mean square error (RMSE) to understand under which conditions, and to which extent the presence of duplicates is problematic.

The results showed that duplicates may bias regression estimates when duplicate records are located in specific parts of the distribution. In other words, the bias was null when the duplicates were randomly drawn from the overall distribution of the dependent variable (Variant *i*). Interestingly, duplicating "typical" cases (Variant *ii*) was just as problematic as duplicating "deviant" cases (Variants *iii* and *iv*). In our simulation the bias was rather low: it reached the highest value of about 7% when the data contained 31 quintuplets. Overall, the bias increased with the share of duplicates in the data, and it was higher for quintuplets than for doublets.

The presence of duplicates in the data affected also the standard errors, and therefore the confidence intervals. Similarly as in the case of the percentage bias, duplicates randomly chosen from the overall distribution (Variant *i*) did not affect standard errors. Duplicating "typical" cases (Variant *ii*) biased the standard errors downwards, thus increasing the statistical power of the estimates. On the contrary, the presence of "deviant" cases (Variants *iii* and *iv*) biased the standard errors upwards, thus producing less precise estimates. The bias of standard errors was overall low (up to maximum 3%), and it was higher when more duplicates were present in the data.

The presence of duplicates also affected the risk of obtaining biased estimates. We considered as biased the coefficients that departed by at least 0.5 standard errors from the true value. The risk of obtaining biased estimates increased with the share of duplicates in the data, reaching the values between 44%–59% (depending on the Variant) when 31 quintuplets were present in the data. The risk was also higher when the duplicates were located on the ties of the distribution (Variants *iii* and *iv*), and it was the lowest when the duplicates were located in the center of the distribution (Variant *ii*). Also the pattern of duplicates mattered, with quintuplets being more problematic than doublets.

The above results are interesting in the light of previous studies which discussed the possible consequences of duplicates for regression estimates. We found no evidence of *attenuation bias*, which suggests that duplicates do not introduce random noise in the data. Moreover, we did not find any

bias when duplicates were located randomly on the overall distribution. On the other hand, if the duplicates were located in a specific part of the distribution, the bias was systematic. Moreover, we found that duplicates increased the *statistical power* of estimates if the duplicated cases were located in the center of the distribution. On the contrary, when duplicates were located on the ties of the distribution, they biased the confidence intervals upwards. We also found that duplication of *"typical"* cases is as problematic as duplication of *"deviant"* cases. It may be even considered more problematic because the bias produced by "typical" duplicates is accompanied by narrower confidence intervals, i.e. higher statistical significance. On the other hand, biased coefficients produced by "deviant" duplicates are accompanied by broader confidence intervals.

The number and patterns of duplicate records used in this analysis are consistent with those identified by Slomczynski et al. (2017), 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 risk of obtaining biased estimates. Thus, researchers who use data with duplicate records risk to reach misleading conclusions.

The second goal of our analysis was to investigate the efficacy of four solutions to reduce the effect of duplicates on estimation results. They included: (*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.

The techniques that performed the best are solutions *d* and *e*, which basically reduced the bias of the coefficient to zero. They also performed well in reducing the risk of obtaining biased estimates. The downside is that these solutions biased upwards the estimated standard errors. Hence, although dropping the superfluous duplicates or weighting the observations by the inverse of the duplicates' multiplicity allow to obtain unbiased coefficients, these solutions come at the cost of decreasing the statistical power of estimates.

The solution which performed the worst was flagging duplicates and controlling for them in the estimation. It produced coefficients' estimates that were more biased than those obtained in the naive estimation. Additionally, it systematically underestimated the standard errors. This is a particularly worrisome combination because biased coefficients were associated to a higher statistical confidence.

Hence, the second conclusion from our study 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 percentage bias, in reducing the risk of obtaining biased estimates, and minimizing the RMSE. Unfortunately, they are associated to larger standard errors, and therefore to lower statistical power. Flagging duplicates and controlling for them is consistently a worst solution, and in some cases (especially for Scenario 2) it produces a higher bias, narrower confidence intervals, 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 considered an ideal case in which the model used by researchers perfectly fitted the relationship in the data, i.e. all relevant predictors were included in the model. This is an unusual situation in social research. Second, our study did 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 considered a situation without unmodeled interactions among variables. Third, in our model the records which were 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 the most difficult to reach by interviewers. Plausibly, the omitted variables, the heterogeneity of the population, and the non-random choice of the interviews replaced by the duplicates exacerbate the impact of duplicates on regression coefficients. This suggests that our estimates of the effect of duplicates on percentage bias, standard errors, 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. Addressing these limitations is 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 for more refined statistical tools to minimize the consequent bias of coefficients and standard errors, the risk of obtaining biased estimates, and the efficiency loss.

Acknowledgements

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.

References

American Statistical Association. (2004). Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects. *Survey Research. Newsletter from the Survey Research Laboratory, College of Urban Planning and Public Affairs, University of Illinois at Chicago. 35*(1), 1–5.

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., & Verykios, V. S. (2007). Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1–16.

European Social Survey. (2015). *European Social Survey round 6.* 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. & Ranchhod, V. (2013). *Genuine fakes: The prevalence and implications of fieldworker fraud in a large South African survey*. Working Paper 115 of the Southern Africa Labour and Development Research Unit, University of Cape Town, 115.

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

Hassanzadeh, O. & 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., & Mushtaq, A. (2015). Curbstoning and beyond: confronting data fabrication in survey research. *Statistical Journal of the IAOS*, *31*(3), 413–422.

Kuriakose, N. & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, *32*(3), 283–291.

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

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

Schräpler, J.-.-.-P. & 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., & Newbrough, J. (1988). Interviewer falsification in Census Bureau surveys. In *Proceedings of the American Statistical Association (Survey Research Methods Section)* (pp. 491–496).

Slomczynski, K. M., Powałko, P., & Krauze, T. (2017). Non-unique records in International Survey Projects: The need for extending data quality control. *Survey Research Methods*, *11*(1), 1–16. doi:doi:10.18148/srm/2017.v11i1.6557

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.

Appendix A
Descriptive statistics for the simulated data sets.
(see table A1 below)

Appendix B
Percentage bias for the remaining coefficients
(see tables B1–B3 below)

Appendix C
Standard errors for the remaining coefficients
(see tables C1–C3 below)

Appendix D
Root mean square error for the remaining coefficients
(see tables D1–D3 below)

Table A1
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
0	x	48.04	16.86	−14.13	99.64	1500	0
0	z	4.916	2.402	−3.839	13.99	1500	0
0	t	40.03	13.35	−5.743	90.41	1500	0
Scenario 1							
1 sextuplet	y	5.212	2.587	−3.878	14.02	1500	0
1 sextuplet	x	47.99	16.83	−14.13	99.64	1500	0
1 sextuplet	z	4.911	2.400	−3.839	13.99	1500	0
1 sextuplet	t	40.01	13.33	−5.743	90.41	1500	0
1 sextuplet	duplicates (flag)	0.00400	0.0631	0	1	1500	0
1 octuplet	y	5.225	2.588	−3.878	14.02	1500	0
1 octuplet	x	47.96	16.89	−14.13	99.64	1500	0
1 octuplet	z	4.930	2.403	−3.839	13.99	1500	0
1 octuplet	t	39.90	13.43	−5.743	90.41	1500	0
1 octuplet	duplicates (flag)	0.00533	0.0729	0	1	1500	0
1 decuplet	y	5.187	2.595	−3.878	14.02	1500	0
1 decuplet	x	47.93	16.87	−14.13	99.64	1500	0
1 decuplet	z	4.909	2.393	−3.839	13.99	1500	0
1 decuplet	t	39.98	13.28	−5.743	90.41	1500	0
1 decuplet	duplicates (flag)	0.00667	0.0814	0	1	1500	0
Scenario 2							
16 doublets	y	5.217	2.582	−3.878	14.02	1500	0
16 doublets	x	48.06	16.92	−14.13	99.64	1500	0
16 doublets	z	4.933	2.409	−3.839	13.99	1500	0
16 doublets	t	40.00	13.32	−5.743	90.41	1500	0
16 doublets	duplicates (flag)	0.0213	0.145	0	1	1500	0
40 doublets	y	5.219	2.599	−3.878	14.02	1500	0
40 doublets	x	48.18	16.81	−14.13	99.64	1500	0
40 doublets	z	4.929	2.410	−3.839	13.99	1500	0
40 doublets	t	39.94	13.44	−5.743	90.41	1500	0
40 doublets	duplicates (flag)	0.0533	0.225	0	1	1500	0
79 doublets	y	5.227	2.582	−3.878	14.02	1500	0
79 doublets	x	47.99	16.94	−14.13	99.64	1500	0
79 doublets	z	4.896	2.404	−3.839	13.99	1500	0
79 doublets	t	40.01	13.42	−5.743	90.41	1500	0
79 doublets	duplicates (flag)	0.105	0.307	0	1	1500	0
Scenario 3							
7 quintuplets	y	5.219	2.584	−3.878	14.02	1500	0
7 quintuplets	x	48.09	16.87	−14.13	99.64	1500	0
7 quintuplets	z	4.932	2.404	−3.839	13.99	1500	0
7 quintuplets	t	39.81	13.47	−5.743	90.41	1500	0
7 quintuplets	duplicates (flag)	0.0233	0.151	0	1	1500	0
16 quintuplets	y	5.240	2.631	−3.878	14.02	1500	0
16 quintuplets	x	48.08	16.71	−14.13	99.64	1500	0
16 quintuplets	z	4.918	2.453	−3.839	13.99	1500	0
16 quintuplets	t	39.91	13.44	−5.743	90.41	1500	0
16 quintuplets	duplicates (flag)	0.0533	0.225	0	1	1500	0
31 quintuplets	y	5.198	2.598	−3.878	14.02	1500	0
31 quintuplets	x	48.15	16.61	−14.13	97.58	1500	0
31 quintuplets	z	4.941	2.458	−3.839	13.99	1500	0
31 quintuplets	t	39.71	13.39	−5.743	90.41	1500	0
31 quintuplets	duplicates (flag)	0.103	0.304	0	1	1500	0

Table B1
Percentage bias of the β_z coefficient (expressed as a percentage of β_z)

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	0.0	0.0	0.0	0.0	0.0
1 octuplet	0.0	−0.0	−0.0	−0.0	−0.0
1 decuplet	−0.1	−0.0	−0.0	−0.0	−0.0
Variant ii: "typical"					
1 sextuplet	−0.3	0.1	0.1	0.0	0.0
1 octuplet	−0.3	0.0	0.0	−0.0	−0.0
1 decuplet	−0.5	0.0	0.0	−0.0	−0.0
Variant iii: "deviant"					
1 sextuplet	0.3	−0.0	−0.0	0.0	0.0
1 octuplet	0.2	−0.1	−0.1	−0.0	−0.0
1 decuplet	0.5	−0.0	−0.0	0.0	0.0
Variant iv: "deviant"					
1 sextuplet	0.3	−0.0	−0.0	0.0	0.0
1 octuplet	0.4	−0.1	−0.1	−0.0	−0.0
1 decuplet	0.6	−0.0	−0.0	0.0	0.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	0.0	−0.1	0.0	−0.0	−0.0
40 doublets	−0.0	0.0	−0.0	0.0	0.0
79 doublets	0.0	0.0	0.0	0.0	0.0
Variant ii: "typical"					
16 doublets	−0.9	0.9	−0.8	0.0	0.0
40 doublets	−2.3	2.3	−2.2	−0.1	−0.1
79 doublets	−4.7	4.6	−4.6	−0.3	−0.3
Variant iii: "deviant"					
16 doublets	0.7	−0.7	−1.5	−0.0	−0.0
40 doublets	1.9	−1.9	−4.0	0.2	0.2
79 doublets	3.5	−4.1	−8.2	0.2	0.2
Variant iv: "deviant"					
16 doublets	1.1	−1.1	−2.1	0.0	0.0
40 doublets	2.7	−2.8	−5.0	0.2	0.2
79 doublets	5.3	−6.0	−10.3	0.5	0.5
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	0.2	0.1	0.1	0.1	0.1
16 quintuplets	−0.1	−0.2	−0.1	−0.2	−0.2
31 quintuplets	0.2	0.1	0.2	0.1	0.1
Variant ii: "typical"					
7 quintuplets	−1.5	0.4	−1.2	0.0	0.0
16 quintuplets	−3.6	0.8	−3.3	−0.1	−0.1
31 quintuplets	−6.8	1.9	−6.6	−0.0	−0.0
Variant iii: "deviant"					
7 quintuplets	1.2	−0.3	−1.1	0.0	0.0
16 quintuplets	2.6	−0.8	−2.9	−0.1	−0.1
31 quintuplets	4.8	−1.5	−5.5	0.1	0.1
Variant iv: "deviant"					
7 quintuplets	2.0	−0.5	−1.3	0.1	0.1
16 quintuplets	4.5	−1.2	−3.2	0.0	0.0
31 quintuplets	7.0	−2.1	−6.6	0.1	0.1

Table B2
Percentage bias of the β_t coefficient (expressed as a percentage of β_t)

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	−0.1	−0.0	−0.0	−0.0	−0.0
1 octuplet	0.0	−0.0	−0.0	−0.0	−0.0
1 decuplet	0.1	0.0	0.0	0.0	0.0
Variant ii: "typical"					
1 sextuplet	−0.2	0.1	0.1	0.1	0.1
1 octuplet	−0.3	0.1	0.1	0.0	0.0
1 decuplet	−0.6	0.1	0.1	0.0	0.0
Variant iii: "deviant"					
1 sextuplet	0.2	0.0	0.0	0.0	0.0
1 octuplet	0.3	−0.0	−0.0	0.0	0.0
1 decuplet	0.8	−0.1	−0.1	−0.0	−0.0
Variant iv: "deviant"					
1 sextuplet	0.4	−0.0	−0.0	0.1	0.1
1 octuplet	0.4	−0.0	−0.0	0.0	0.0
1 decuplet	0.4	−0.1	−0.1	−0.0	−0.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	−0.1	0.0	−0.1	−0.0	−0.0
40 doublets	0.2	0.1	0.2	0.1	0.1
79 doublets	−0.2	−0.2	−0.1	−0.2	−0.2
Variant ii: "typical"					
16 doublets	−0.9	1.0	−0.8	−0.0	−0.0
40 doublets	−2.4	2.2	−2.3	−0.2	−0.2
79 doublets	−4.8	4.4	−4.7	−0.4	−0.4
Variant iii: "deviant"					
16 doublets	0.9	−0.7	−1.9	0.1	0.1
40 doublets	1.9	−2.2	−5.2	0.0	0.0
79 doublets	3.7	−4.6	−10.7	0.3	0.3
Variant iv: "deviant"					
16 doublets	0.9	−1.0	−2.1	−0.0	−0.0
40 doublets	2.6	−3.1	−5.8	−0.0	−0.0
79 doublets	5.0	−6.0	−11.0	0.3	0.3
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	0.0	−0.0	0.0	−0.0	−0.0
16 quintuplets	−0.1	−0.0	−0.1	−0.0	−0.0
31 quintuplets	0.4	0.0	0.3	0.1	0.1
Variant ii: "typical"					
7 quintuplets	−1.6	0.3	−1.3	−0.1	−0.1
16 quintuplets	−3.8	1.0	−3.6	−0.0	−0.0
31 quintuplets	−7.2	1.6	−7.0	−0.3	−0.3
Variant iii: "deviant"					
7 quintuplets	1.4	−0.3	−1.5	0.1	0.1
16 quintuplets	2.9	−0.8	−3.6	0.0	0.0
31 quintuplets	5.0	−1.6	−7.6	0.1	0.1
Variant iv: "deviant"					
7 quintuplets	1.7	−0.4	−1.6	0.0	0.0
16 quintuplets	4.3	−1.1	−3.7	0.1	0.1
31 quintuplets	7.3	−2.3	−7.6	0.1	0.1

Table B3
Percentage bias of the intercept (expressed as a percentage of the intercept)

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	−0.0	0.0	0.0	0.0	0.0
1 octuplet	−0.0	−0.0	−0.0	−0.0	−0.0
1 decuplet	−0.0	−0.0	−0.0	−0.0	−0.0
Variant ii: "typical"					
1 sextuplet	−0.0	−0.0	−0.0	−0.0	−0.0
1 octuplet	−0.0	0.0	0.0	0.0	0.0
1 decuplet	−0.0	−0.0	−0.0	−0.0	−0.0
Variant iii: "deviant"					
1 sextuplet	−0.1	0.0	0.0	−0.0	−0.0
1 octuplet	−0.2	0.0	0.0	0.0	0.0
1 decuplet	−0.3	0.0	0.0	0.0	0.0
Variant iv: "deviant"					
1 sextuplet	0.2	−0.0	−0.0	−0.0	−0.0
1 octuplet	0.2	−0.0	−0.0	0.0	0.0
1 decuplet	0.3	−0.0	−0.0	0.0	0.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	0.0	−0.0	−0.0	−0.0	−0.0
40 doublets	−0.0	−0.0	−0.0	−0.0	−0.0
79 doublets	0.0	0.1	0.0	0.0	0.0
Variant ii: "typical"					
16 doublets	−0.0	0.0	−0.0	−0.0	−0.0
40 doublets	−0.0	0.1	−0.0	0.0	0.0
79 doublets	−0.1	0.2	−0.0	0.0	0.0
Variant iii: "deviant"					
16 doublets	−0.5	0.5	0.1	−0.0	−0.0
40 doublets	−1.2	1.2	0.3	−0.1	−0.1
79 doublets	−2.5	2.4	0.7	−0.1	−0.1
Variant iv: "deviant"					
16 doublets	0.5	−0.5	−0.8	0.0	0.0
40 doublets	1.3	−1.3	−2.0	0.0	0.0
79 doublets	2.5	−2.7	−4.1	0.1	0.1
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	−0.0	−0.0	0.0	−0.0	−0.0
16 quintuplets	0.0	0.0	0.0	0.0	0.0
31 quintuplets	−0.1	−0.0	−0.1	−0.0	−0.0
Variant ii: "typical"					
7 quintuplets	−0.1	0.0	−0.0	−0.0	−0.0
16 quintuplets	−0.1	0.0	−0.1	−0.0	−0.0
31 quintuplets	−0.2	0.1	−0.1	0.0	0.0
Variant iii: "deviant"					
7 quintuplets	−0.9	0.2	−0.2	−0.0	−0.0
16 quintuplets	−1.9	0.5	−0.4	−0.0	−0.0
31 quintuplets	−3.8	0.9	−0.8	−0.1	−0.1
Variant iv: "deviant"					
7 quintuplets	0.9	−0.2	−0.5	−0.0	−0.0
16 quintuplets	1.9	−0.5	−1.2	0.0	0.0
31 quintuplets	4.0	−1.0	−2.2	0.1	0.1

Table C1
Average standard error of β_z coefficient (expressed as a percentage of the true standard error of β_z).

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	100.0	100.2	100.0	100.2	101.9
1 octuplet	100.0	100.3	100.0	100.2	102.0
1 decuplet	100.0	100.3	100.0	100.3	102.0
Variant ii: "typical"					
1 sextuplet	99.9	100.2	100.1	100.2	101.9
1 octuplet	99.8	100.3	100.1	100.2	102.0
1 decuplet	99.8	100.4	100.1	100.3	102.1
Variant iii: "deviant"					
1 sextuplet	100.1	100.2	100.0	100.2	101.9
1 octuplet	100.2	100.2	100.0	100.2	101.9
1 decuplet	100.2	100.3	100.0	100.3	102.0
Variant iv: "deviant"					
1 sextuplet	100.1	100.2	100.0	100.2	101.9
1 octuplet	100.2	100.2	100.0	100.2	101.9
1 decuplet	100.3	100.3	100.0	100.3	102.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	100.0	101.1	100.0	100.5	102.0
40 doublets	100.0	102.8	100.0	101.4	102.4
79 doublets	100.0	105.7	100.0	102.7	103.1
Variant ii: "typical"					
16 doublets	99.6	101.5	99.7	100.5	102.2
40 doublets	98.9	103.9	99.0	101.3	102.9
79 doublets	97.7	108.1	97.8	102.6	104.0
Variant iii: "deviant"					
16 doublets	100.4	100.7	99.0	100.5	101.9
40 doublets	101.0	101.7	97.3	101.4	102.0
79 doublets	101.8	103.3	94.4	102.9	102.3
Variant iv: "deviant"					
16 doublets	100.4	100.6	99.0	100.6	101.8
40 doublets	101.1	101.5	97.3	101.4	101.9
79 doublets	102.1	103.0	94.4	102.9	102.0
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	100.0	101.2	100.0	100.9	102.5
16 quintuplets	100.0	102.8	100.0	102.2	103.5
31 quintuplets	100.0	105.6	100.0	104.4	105.3
Variant ii: "typical"					
7 quintuplets	99.3	101.4	99.4	100.9	102.7
16 quintuplets	98.3	103.2	98.4	102.2	103.8
31 quintuplets	96.6	106.5	96.8	104.3	105.9
Variant iii: "deviant"					
7 quintuplets	100.7	101.0	99.2	101.0	102.4
16 quintuplets	101.5	102.4	98.1	102.2	103.3
31 quintuplets	102.7	104.7	96.1	104.5	104.7
Variant iv: "deviant"					
7 quintuplets	100.8	101.0	99.3	101.0	102.4
16 quintuplets	101.7	102.3	98.1	102.2	103.2
31 quintuplets	103.0	104.7	96.2	104.5	104.6

Table C2
Average standard error of β_t coefficient (expressed as a percentage of the true standard error of β_t).

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	100.0	100.2	100.0	100.2	106.6
1 octuplet	100.0	100.3	100.0	100.2	106.6
1 decuplet	100.0	100.3	100.0	100.3	106.7
Variant ii: "typical"					
1 sextuplet	99.9	100.2	100.1	100.2	106.6
1 octuplet	99.8	100.3	100.1	100.2	106.7
1 decuplet	99.8	100.4	100.1	100.3	106.7
Variant iii: "deviant"					
1 sextuplet	100.1	100.2	100.0	100.2	106.5
1 octuplet	100.2	100.2	100.0	100.2	106.6
1 decuplet	100.2	100.3	100.0	100.3	106.7
Variant iv: "deviant"					
1 sextuplet	100.1	100.2	100.0	100.2	106.5
1 octuplet	100.2	100.2	100.0	100.2	106.6
1 decuplet	100.2	100.3	100.0	100.3	106.7
Scenario 2					
Variant i: "unconstrained"					
16 doublets	100.0	101.1	100.0	100.5	106.7
40 doublets	100.0	102.8	100.0	101.4	107.1
79 doublets	100.0	105.8	100.1	102.8	107.8
Variant ii: "typical"					
16 doublets	99.6	101.5	99.7	100.5	106.9
40 doublets	98.9	103.9	99.0	101.3	107.6
79 doublets	97.8	108.0	97.9	102.6	108.8
Variant iii: "deviant"					
16 doublets	100.4	100.6	99.0	100.5	106.5
40 doublets	101.0	101.6	97.3	101.4	106.5
79 doublets	101.9	103.2	94.5	102.9	106.8
Variant iv: "deviant"					
16 doublets	100.4	100.7	98.9	100.5	106.5
40 doublets	101.0	101.6	97.2	101.4	106.6
79 doublets	101.9	103.2	94.1	102.9	106.8
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	100.0	101.2	100.0	101.0	107.2
16 quintuplets	100.0	102.8	100.0	102.2	108.2
31 quintuplets	100.0	105.6	100.1	104.4	110.1
Variant ii: "typical"					
7 quintuplets	99.3	101.4	99.4	100.9	107.4
16 quintuplets	98.3	103.2	98.4	102.2	108.6
31 quintuplets	96.7	106.5	96.9	104.4	110.7
Variant iii: "deviant"					
7 quintuplets	100.7	101.0	99.3	101.0	107.1
16 quintuplets	101.5	102.3	98.1	102.2	107.9
31 quintuplets	102.7	104.7	96.1	104.5	109.4
Variant iv: "deviant"					
7 quintuplets	100.7	101.0	99.2	101.0	107.1
16 quintuplets	101.5	102.4	97.9	102.2	108.0
31 quintuplets	102.6	104.7	95.7	104.5	109.5

Table C3
Average standard error of the intercept (expressed as a percentage of the true standard error of the intercept).

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Variant i: "unconstrained"					
1 sextuplet	100.0	100.2	100.0	100.2	103.9
1 octuplet	100.0	100.3	100.0	100.2	104.0
1 decuplet	100.0	100.3	100.0	100.3	104.1
Variant ii: "typical"					
1 sextuplet	99.9	100.2	100.1	100.2	104.0
1 octuplet	99.8	100.3	100.1	100.2	104.0
1 decuplet	99.8	100.4	100.1	100.3	104.1
Variant iii: "deviant"					
1 sextuplet	100.1	100.2	100.0	100.2	103.9
1 octuplet	100.2	100.2	100.0	100.2	104.0
1 decuplet	100.3	100.3	100.0	100.3	104.0
Variant iv: "deviant"					
1 sextuplet	100.1	100.2	100.0	100.2	103.9
1 octuplet	100.2	100.2	100.0	100.2	104.0
1 decuplet	100.2	100.3	100.0	100.3	104.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	100.0	101.1	100.1	100.5	104.1
40 doublets	100.0	102.8	100.2	101.4	104.5
79 doublets	100.0	105.8	100.3	102.8	105.2
Variant ii: "typical"					
16 doublets	99.6	101.5	99.7	100.5	104.3
40 doublets	98.9	103.9	99.2	101.3	105.0
79 doublets	97.8	108.0	98.2	102.6	106.1
Variant iii: "deviant"					
16 doublets	100.5	100.6	99.0	100.5	103.9
40 doublets	101.1	101.5	97.4	101.4	104.0
79 doublets	102.1	103.0	94.6	102.9	104.3
Variant iv: "deviant"					
16 doublets	100.4	100.7	99.0	100.5	103.9
40 doublets	101.0	101.7	97.2	101.4	104.0
79 doublets	101.8	103.3	94.3	102.9	104.1
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	100.0	101.2	100.0	100.9	104.6
16 quintuplets	100.0	102.8	100.1	102.2	105.6
31 quintuplets	100.0	105.6	100.3	104.4	107.4
Variant ii: "typical"					
7 quintuplets	99.3	101.4	99.5	100.9	104.7
16 quintuplets	98.3	103.2	98.6	102.2	105.9
31 quintuplets	96.7	106.5	97.1	104.4	108.0
Variant iii: "deviant"					
7 quintuplets	100.8	101.0	99.3	101.0	104.4
16 quintuplets	101.7	102.3	98.2	102.2	105.3
31 quintuplets	103.0	104.6	96.3	104.5	106.8
Variant iv: "deviant"					
7 quintuplets	100.7	101.0	99.2	101.0	104.4
16 quintuplets	101.5	102.4	97.9	102.2	105.2
31 quintuplets	102.5	104.8	95.8	104.5	106.7

Table D1
Normalized RMSE for the β_z coefficient

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	14.8	14.8	14.8	14.8	15.0
1 octuplet	14.8	14.8	14.8	14.8	15.1
1 decuplet	14.8	14.8	14.8	14.8	15.1
Variant ii: "typical"					
1 sextuplet	14.7	14.8	14.8	14.8	15.0
1 octuplet	14.7	14.8	14.8	14.8	15.1
1 decuplet	14.7	14.8	14.8	14.8	15.1
Variant iii: "deviant"					
1 sextuplet	14.8	14.8	14.8	14.8	15.0
1 octuplet	14.8	14.8	14.8	14.8	15.1
1 decuplet	14.8	14.8	14.8	14.8	15.1
Variant iv: "deviant"					
1 sextuplet	14.8	14.8	14.8	14.8	15.0
1 octuplet	14.8	14.8	14.8	14.8	15.1
1 decuplet	14.8	14.8	14.8	14.8	15.1
Scenario 2					
Variant i: "unconstrained"					
16 doublets	14.8	14.9	14.8	14.8	15.1
40 doublets	14.8	15.2	14.8	15.0	15.1
79 doublets	14.8	15.6	14.8	15.2	15.2
Variant ii: "typical"					
16 doublets	14.7	15.0	14.7	14.8	15.1
40 doublets	14.8	15.5	14.8	15.0	15.2
79 doublets	15.2	16.6	15.2	15.2	15.4
Variant iii: "deviant"					
16 doublets	14.8	14.9	14.7	14.8	15.0
40 doublets	15.0	15.1	14.9	15.0	15.1
79 doublets	15.4	15.8	16.2	15.2	15.1
Variant iv: "deviant"					
16 doublets	14.9	14.9	14.8	14.8	15.0
40 doublets	15.2	15.2	15.2	15.0	15.0
79 doublets	16.0	16.3	17.3	15.2	15.1
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	14.8	14.9	14.8	14.9	15.1
16 quintuplets	14.8	15.2	14.8	15.1	15.3
31 quintuplets	14.8	15.6	14.8	15.4	15.5
Variant ii: "typical"					
7 quintuplets	14.7	15.0	14.7	14.9	15.2
16 quintuplets	14.9	15.3	14.9	15.1	15.3
31 quintuplets	15.8	15.8	15.7	15.4	15.6
Variant iii: "deviant"					
7 quintuplets	14.9	14.9	14.7	14.9	15.1
16 quintuplets	15.2	15.1	14.8	15.1	15.2
31 quintuplets	15.9	15.5	15.2	15.4	15.5
Variant iv: "deviant"					
7 quintuplets	15.0	14.9	14.7	14.9	15.1
16 quintuplets	15.7	15.2	14.8	15.1	15.2
31 quintuplets	16.7	15.6	15.7	15.4	15.4

Table D2
Normalized RMSE for the β_t coefficient

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Variant i: "unconstrained"					
1 sextuplet	24.3	24.3	24.3	24.3	25.9
1 octuplet	24.3	24.4	24.3	24.4	25.9
1 decuplet	24.3	24.4	24.3	24.4	25.9
Variant ii: "typical"					
1 sextuplet	24.3	24.4	24.3	24.3	25.9
1 octuplet	24.3	24.4	24.3	24.4	25.9
1 decuplet	24.2	24.4	24.3	24.4	25.9
Variant iii: "deviant"					
1 sextuplet	24.3	24.3	24.3	24.3	25.9
1 octuplet	24.3	24.4	24.3	24.4	25.9
1 decuplet	24.4	24.4	24.3	24.4	25.9
Variant iv: "deviant"					
1 sextuplet	24.3	24.3	24.3	24.3	25.9
1 octuplet	24.3	24.4	24.3	24.4	25.9
1 decuplet	24.4	24.4	24.3	24.4	25.9
Scenario 2					
Variant i: "unconstrained"					
16 doublets	24.3	24.6	24.3	24.4	25.9
40 doublets	24.3	25.0	24.3	24.6	26.0
79 doublets	24.3	25.7	24.3	25.0	26.2
Variant ii: "typical"					
16 doublets	24.2	24.7	24.2	24.4	26.0
40 doublets	24.2	25.3	24.2	24.6	26.1
79 doublets	24.2	26.6	24.2	24.9	26.4
Variant iii: "deviant"					
16 doublets	24.4	24.5	24.1	24.4	25.9
40 doublets	24.6	24.8	24.2	24.6	25.9
79 doublets	25.0	25.5	25.3	25.0	25.9
Variant iv: "deviant"					
16 doublets	24.4	24.5	24.1	24.4	25.9
40 doublets	24.7	24.9	24.3	24.6	25.9
79 doublets	25.3	25.8	25.4	25.0	26.0
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	24.3	24.6	24.3	24.5	26.1
16 quintuplets	24.3	25.0	24.3	24.8	26.3
31 quintuplets	24.3	25.7	24.3	25.4	26.8
Variant ii: "typical"					
7 quintuplets	24.2	24.6	24.2	24.5	26.1
16 quintuplets	24.2	25.1	24.2	24.8	26.4
31 quintuplets	24.6	25.9	24.5	25.4	26.9
Variant iii: "deviant"					
7 quintuplets	24.5	24.5	24.2	24.5	26.0
16 quintuplets	24.8	24.9	24.1	24.8	26.2
31 quintuplets	25.5	25.5	24.6	25.4	26.6
Variant iv: "deviant"					
7 quintuplets	24.5	24.5	24.2	24.5	26.0
16 quintuplets	25.0	24.9	24.1	24.8	26.2
31 quintuplets	26.0	25.5	24.5	25.4	26.6

Table D3
Normalized RMSE for the intercept

	Solution				
	(a) "Naive" estimation	(b) Drop all	(c) Flag and control	(d) Drop superfluous	(e) Weighted regression
Scenario 1					
Variant i: "unconstrained"					
1 sextuplet	5.8	5.8	5.8	5.8	6.0
1 octuplet	5.8	5.8	5.8	5.8	6.0
1 decuplet	5.8	5.8	5.8	5.8	6.0
Variant ii: "typical"					
1 sextuplet	5.7	5.8	5.8	5.8	6.0
1 octuplet	5.7	5.8	5.8	5.8	6.0
1 decuplet	5.7	5.8	5.8	5.8	6.0
Variant iii: "deviant"					
1 sextuplet	5.8	5.8	5.8	5.8	6.0
1 octuplet	5.8	5.8	5.8	5.8	6.0
1 decuplet	5.8	5.8	5.8	5.8	6.0
Variant iv: "deviant"					
1 sextuplet	5.8	5.8	5.8	5.8	6.0
1 octuplet	5.8	5.8	5.8	5.8	6.0
1 decuplet	5.8	5.8	5.8	5.8	6.0
Scenario 2					
Variant i: "unconstrained"					
16 doublets	5.8	5.8	5.8	5.8	6.0
40 doublets	5.7	5.9	5.8	5.8	6.0
79 doublets	5.8	6.1	5.8	5.9	6.0
Variant ii: "typical"					
16 doublets	5.7	5.8	5.7	5.8	6.0
40 doublets	5.7	6.0	5.7	5.8	6.0
79 doublets	5.6	6.2	5.6	5.9	6.1
Variant iii: "deviant"					
16 doublets	5.8	5.8	5.7	5.8	6.0
40 doublets	5.9	6.0	5.6	5.8	6.0
79 doublets	6.4	6.4	5.5	5.9	6.0
Variant iv: "deviant"					
16 doublets	5.8	5.8	5.7	5.8	6.0
40 doublets	5.9	6.0	5.9	5.8	6.0
79 doublets	6.4	6.5	6.8	5.9	6.0
Scenario 3					
Variant i: "unconstrained"					
7 quintuplets	5.7	5.8	5.8	5.8	6.0
16 quintuplets	5.7	5.9	5.8	5.9	6.1
31 quintuplets	5.8	6.1	5.8	6.0	6.2
Variant ii: "typical"					
7 quintuplets	5.7	5.8	5.7	5.8	6.0
16 quintuplets	5.7	5.9	5.7	5.9	6.1
31 quintuplets	5.6	6.1	5.6	6.0	6.2
Variant iii: "deviant"					
7 quintuplets	5.9	5.8	5.7	5.8	6.0
16 quintuplets	6.2	5.9	5.7	5.9	6.1
31 quintuplets	7.1	6.1	5.6	6.0	6.1
Variant iv: "deviant"					
7 quintuplets	5.9	5.8	5.7	5.8	6.0
16 quintuplets	6.1	5.9	5.8	5.9	6.1
31 quintuplets	7.1	6.1	5.9	6.0	6.1