Schmid, Matthias; Schneeweiss, Hans

Working Paper

The effect of microaggregation procedures on the estimation of linear models: a simulation study

Discussion Paper, No. 443

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Schmid, Matthias; Schneeweiss, Hans (2005) : The effect of microaggregation procedures on the estimation of linear models: a simulation study, Discussion Paper, No. 443, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München, https://doi.org/10.5282/ubm/epub.1812

This Version is available at:
https://hdl.handle.net/10419/31046

The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study

Matthias Schmid[1] **and Hans Schneeweiss**[2]

[1] Department of Statistics, University of Munich
Ludwigstr. 33, D-80539 München, Germany
Email: matthias.schmid@stat.uni-muenchen.de

[2] Department of Statistics, University of Munich
Akademiestr. 1, D-80799 München, Germany
Email: hans.schneeweiss@stat.uni-muenchen.de

Abstract

Microaggregation is a set of procedures that distort empirical data in order to guarantee the factual anonymity of the data. At the same time the information content of data sets should not be reduced too much and should still be useful for scientific research. This paper investigates the effect of microaggregation on the estimation of a linear regression by ordinary least squares. It studies, by way of an extensive simulation experiment, the bias of the slope parameter estimator induced by various microaggregation techniques. Some microaggregation procedures lead to consistent estimates while others imply an asymptotic bias for the estimator.

Keywords: microaggregation, disclosure control, simple linear model, bias, consistency

1 Introduction

A problem statistical offices are increasingly faced with is providing sufficient information to scientists while at the same time having to maintain confidentiality required by data protection laws. To handle this trade-off (which

is commonly referred to as the *statistical disclosure control problem*), the information content of released microdata sets is often reduced by means of masking procedures. However, while reducing the disclosure risk of a data file, masking procedures also affect the results of statistical analyses.

One of the most promising masking techniques is microaggregation (Defays/Anwar 1998, Domingo-Ferrer/Mateo-Sanz 2002), a procedure for continuous data which has been widely discussed over the last years. The main idea behind microaggregation is to group the observations in a data set and replace the original data values with their corresponding group means. To reduce the information loss imposed by microaggregation, it is considered advisable to group only those data values which are similar in terms of a similarity criterion. The various types of microaggregation techniques mainly differ in the similarity criterion that is used to form the groups.

While the disclosure risk of anonymized data sets has been subject to intensive research over the last years (Elliot 2001, Willenborg/de Waal 2001, Yancey et al. 2002), the impact of microaggregation on statistical analyses is still widely unexplored. Empirical studies based on the analysis of selected data sets include Mateo-Sanz/Domingo-Ferrer (1998), Domingo-Ferrer/Mateo-Sanz (2001), and Domingo-Ferrer/Torra (2001). However, while providing an important insight into the effects of microaggregation on statistical analyses, the results of these studies may depend on various (unknown) characteristics of the data sets and on the uncertainty whether the statistical models are correctly specified. In this paper, we instead focus on simulated data sets. Thus we are able to control the data structure and the model specification and can concentrate on a study of the microaggregation effect solely.

The effect of microaggregation on statistical analyses depends on the type of analysis carried out by the researcher. It is therefore necessary to study all forms of models to be considered for statistical analysis. In the following, we restrict our investigation to the estimation of a simple linear regression model. This is done by means of a systematic simulation study. Our main interest is in the potential bias of the naive estimator of the slope parameter. It turns out that all aggregation methods considered in this paper lead to a bias, at least for small sample sizes. In some cases the bias persists asymptotically while in other cases it decreases to zero with growing sample size, thus giving rise to consistent estimates.

In Section 2 we start with a description of the various microaggregation procedures used for masking data. Section 3 contains a systematic simulation study of the effects of microaggregation on the estimation of a simple linear model. Moreover, we consider in detail the microaggregation techniques which induce an asymptotic bias in the estimated slope parameter. In Section 4, we outline the effects of microaggregation on the estimation of a multiple linear regression model. Section 5 contains a concluding summary.

2 Microaggregation Techniques

As stated in the introduction, there are various types of microaggregation methods which mainly differ in the similarity criterion used for grouping the data. For our simulation study, we decided to chose five of the most commonly applied microaggregation techniques, namely

1. Microaggregation using a leading variable (Paass/Wauschkuhn 1985): The data values are sorted with respect to one variable in the data set (the so-called leading variable). Groups are then formed by data records having similar values for the leading variable. The group size (or "aggregation level") is kept fixed for every group. In a simple linear model, the leading variable can either be the regressor or the dependent variable. Feige/Watts (1972) have shown that if the regressor is used as the leading variable, linear model estimates based on the aggregated data are unbiased. Concerning microaggregation using the dependent variable as the leading variable, Feige/Watts (1972) hint at the possibility that estimates might show an aggregation bias. However, little is known about this bias. Therefore, in the following sections, we restrict to the case where the dependent variable is the leading variable.

2. Individual ranking (Defays/Anwar 1998): Each variable is microaggregated separately. First, the data set is sorted by the first variable, and the values of this variable are microaggregated. Then, the same procedure is repeated for the second variable and so on. Again, the group size is kept fixed.

3. Microaggregation using principal component analysis: Multivariate data are first projected onto the first principle axis. The projected values then serve as a leading variable as described in 1. The group size is kept fixed.

4. Multivariate microaggregation based on Euclidean distances: This type of microaggregation uses the Euclidean distance to determine the similarity of data records. Before microaggregation, all variables are standardized. Again, a fixed group size is used. For details on the algorithm we refer to Domingo-Ferrer/Mateo-Sanz (2002).

5. Multivariate microaggregation based on Ward hierarchical clustering (Domingo-Ferrer/Mateo-Sanz 2002): A cluster analysis with minimum group size is performed to aggregate the data. Under the constraint that each group must consist of at least A data values, groups are formed by minimizing the within-groups variance. The group size is allowed to vary over the groups. Again, we refer to Domingo-Ferrer/Mateo-Sanz (2002) for an exact description of the algorithm.

3 Simulations

We consider the simple linear model

$$Y = \alpha + \beta X + \epsilon \ . \tag{1}$$

Y denotes the continuous response (or endogenous variable) while X denotes the continuous covariate (or exogenous variable). $(\alpha, \beta)'$ is the corresponding parameter vector. The random error ϵ is independent of X. Moreover, ϵ is assumed to be normally distributed with zero mean and constant variance σ_ϵ^2.

We applied the five microaggregation techniques described in Section 2 to data sets simulated from model (1). We then estimated a simple linear model from the aggregated data (so-called *naive* estimation). Table 1 shows the averages of the naive estimates $\hat{\beta}$ over 1000 replications for selected values of β ($\beta = 0$, $\beta = 0.5$, $\beta = 1$, $\beta = 5$) and various sample sizes ($n = 50$, $n = 100$, $n = 200$, $n = 400$). For symmetry reasons we do not consider negative values of β. The residual variance σ_ϵ^2 was set equal to 0.5, α was set equal to 0. The values of the covariate X were drawn from a standard normal distribution.

For microaggregation using Y as a leading variable (LV), microaggregation using individual ranking (IR), microaggregation using principal component analysis (PCA), and microaggregation using Euclidean distances (Eucl), the group size was set equal to three. For microaggregation based on Ward

		n=50	n=100	n=200	n=400
$\beta = 0$	Eucl	-0.001	0.002	-0.001	-0.001
	IR	0.001	0.002	0.000	0.000
	LV	-0.006	0.002	0.000	0.000
	PCA	-0.002	0.004	0.003	-0.007
	kW	0.001	0.001	0.002	0.001
$\beta = 0.5$	Eucl	0.529	0.513	0.510	0.504
	IR	0.492	0.496	0.499	0.499
	LV	0.769	0.762	0.753	0.751
	PCA	0.614	0.611	0.611	0.609
	kW	0.521	0.511	0.507	0.503
$\beta = 1$	Eucl	1.029	1.018	1.008	1.005
	IR	0.988	0.997	0.998	0.998
	LV	1.163	1.158	1.155	1.155
	PCA	1.084	1.085	1.082	1.082
	kW	1.036	1.021	1.012	1.006
$\beta = 5$	Eucl	5.016	5.011	5.009	5.004
	IR	4.972	4.990	4.995	4.998
	LV	5.031	5.035	5.034	5.033
	PCA	5.030	5.027	5.028	5.028
	kW	5.036	5.034	5.028	5.017

Table 1: Naive estimates of β for various sample sizes ($\sigma_\epsilon = 0.5$)

hierarchical clustering (kW), the minimum group size was set equal to three as well.

From Table 1 we see that if n is small and $\beta > 0$, all estimates show a bias. Interestingly, for LV, PCA, Eucl, and kW, the bias is positive, meaning that the effect of X on the dependent variable Y is overestimated. If IR is used to aggregate the data, the slope parameter is underestimated.

The only exception is the case where $\beta = 0$: In this case, for all values of n, the naive estimates are unbiased. For $\beta > 0$ we see that if IR, Eucl, or kW are used for microaggregation, the bias decreases with n increasing and apparently goes to zero with $n \to \infty$. Thus, IR, Eucl, and kW seem to lead to consistent naive estimates of the slope parameter.

In contrast, estimates show an asymptotic bias if LV and PCA are used for aggregation. This bias depends on the value of β. Comparing LV to PCA, we see that PCA induces a smaller bias than LV, at least for the values of β chosen, but see below.

Tables 2 and 3 show what happens if the residual variance σ_ϵ^2 is increased. Apparently, the bias of the estimates becomes larger as σ_ϵ^2 gets larger. In addition, convergence to the true value of β slows down if IR, Eucl, or kW are applied. Interestingly, if $\sigma_\epsilon = 1.5$ and $\beta = 0.5$, the biases of the naive estimates based on LV and PCA are almost equal. In this case, PCA does not perform better than LV. At the end of this section, we will consider in detail the effect of LV and PCA on the naive estimate of β.

		n=50	n=100	n=200	n=400
$\beta = 0$	Eucl	0.004	-0.002	0.002	0.002
	IR	-0.002	-0.001	0.002	-0.002
	LV	-0.026	-0.015	-0.002	0.000
	PCA	-0.026	-0.005	-0.015	0.026
	kW	0.003	-0.001	0.000	-0.003
$\beta = 0.5$	Eucl	0.535	0.519	0.512	0.507
	IR	0.494	0.493	0.499	0.500
	LV	1.120	1.083	1.081	1.076
	PCA	0.911	0.909	0.896	0.890
	kW	0.555	0.525	0.513	0.507
$\beta = 1$	Eucl	1.044	1.034	1.019	1.009
	IR	0.985	0.997	0.999	1.002
	LV	1.545	1.515	1.515	1.508
	PCA	1.330	1.315	1.309	1.310
	kW	1.084	1.047	1.027	1.014
$\beta = 5$	Eucl	5.042	5.027	5.021	5.010
	IR	4.952	4.986	4.987	4.995
	LV	5.136	5.132	5.136	5.130
	PCA	5.115	5.112	5.110	5.112
	kW	5.135	5.107	5.075	5.004

Table 2: Naive estimates of β for various sample sizes ($\sigma_\epsilon = 1$)

		n=50	n=100	n=200	n=400
$\beta = 0$	Eucl	0.013	-0.005	0.000	0.000
	IR	0.000	0.005	-0.004	0.001
	LV	-0.008	-0.007	0.012	0.013
	PCA	0.050	-0.074	0.001	0.071
	kW	-0.003	0.000	0.005	-0.001
$\beta = 0.5$	Eucl	0.537	0.521	0.511	0.510
	IR	0.491	0.490	0.498	0.497
	LV	1.282	1.254	1.262	1.260
	PCA	1.273	1.273	1.263	1.252
	kW	0.576	0.539	0.519	0.511
$\beta = 1$	Eucl	1.050	1.037	1.017	1.012
	IR	0.981	0.991	1.002	0.993
	LV	1.917	1.881	1.874	1.867
	PCA	1.683	1.650	1.638	1.634
	kW	1.146	1.079	1.041	1.025
$\beta = 5$	Eucl	5.087	5.052	5.033	5.009
	IR	4.947	4.970	4.992	4.997
	LV	5.302	5.305	5.293	5.294
	PCA	5.269	5.253	5.247	5.245
	kW	5.276	5.198	5.129	5.074

Table 3: Naive estimates of β for various sample sizes ($\sigma_\epsilon = 1.5$)

		A=3	A=6	A=9	A=12
$\beta = 0$	Eucl	0.002	0.003	-0.002	0.002
	IR	0.002	0.001	0.001	0.000
	LV	-0.002	0.007	0.010	-0.040
	PCA	-0.015	-0.014	0.000	-0.043
	kW	0.000	-0.001	-0.002	0.003
$\beta = 0.5$	Eucl	0.512	0.509	0.511	0.510
	IR	0.499	0.496	0.492	0.494
	LV	1.081	1.536	1.771	1.936
	PCA	0.896	1.024	1.075	1.098
	kW	0.513	0.538	0.562	0.588
$\beta = 1$	Eucl	1.019	1.018	1.018	1.015
	IR	0.999	0.997	0.993	0.981
	LV	1.515	1.731	1.822	1.866
	PCA	1.309	1.410	1.436	1.461
	kW	1.027	1.065	1.109	1.137
$\beta = 5$	Eucl	5.021	5.020	5.022	5.022
	IR	4.987	4.981	4.970	4.960
	LV	5.136	5.168	5.179	5.183
	PCA	5.110	5.141	5.151	5.153
	kW	5.075	5.145	5.173	5.181

Table 4: Naive estimates of β for various group sizes ($\sigma_\epsilon = 1$, $n = 200$)

Table 4 shows the naive estimates of β for various group sizes (here, σ_ϵ was set equal to one and n was set equal to 200). The group size is denoted by A. We see that as A increases and $\beta > 0$, the bias of the naive estimate increases in most cases, particularly if LV or PCA are used for microaggregation. The only exception is the case where Eucl is used to aggregate the data: Here, the group size does not seem to have any effect on the bias of $\hat{\beta}$.

It should be mentioned that the bias of $\hat{\beta}$ induced by kW is not directly comparable to the bias induced by the other microaggregation methods. This is due to the fact that kW allows the group sizes to vary, implying that groups containing more than A records are possible. For instance, if $A = 3$, the average group size of kW estimated from the simulated data sets was about 3.70.

Let us now have a closer look at the nature of the bias of $\hat{\beta}$ if LV or PCA are used to aggregate the data. We estimated the bias of $\hat{\beta}$ for $\beta = 0, 0.005, \ldots, 0.495, 0.5, 0.75, \ldots, 2.75, 3$ and the corresponding negative values. The sample size n was set equal to 400, the group size was set equal to three. As before, the values of X were drawn from a standard normal distribution. The results are shown in Fig. 1 (here, σ_ϵ was set equal to 0.5): Again, we see that if $\beta > 0$, β is overestimated by $\hat{\beta}$. Obviously, if $|\beta|$ is close to zero, PCA induces a larger bias than LV does. For large values of $|\beta|$, PCA performs better. We also see that if $\beta = 0$, estimates are unbiased. As $|\beta| \to \infty$, bias$(\hat{\beta})$ goes to zero. This is a plausible result because using Y as the leading variable is approximately the same as using X as the leading variable when $|\beta|$ is large and σ_ϵ^2 is kept constant (in fact, the correlation between X and Y is close to ± 1 in this case). In the same way, if X and Y are highly correlated, using PCA is almost the same as using X as the leading variable. Now, if X is used as the leading variable, estimates are unbiased (see Feige/Watts 1972), and thus bias$(\hat{\beta})$ should be close to zero if Y is the leading variable and $|\beta|$ is large.

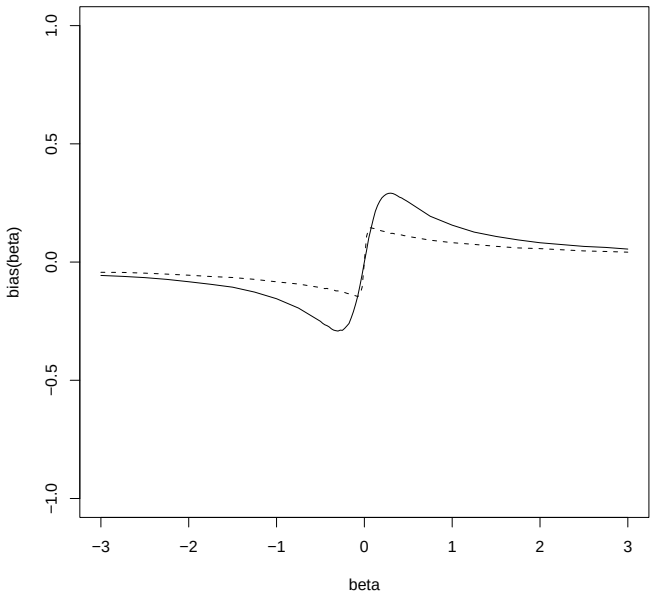

Figure 1: Bias of $\hat{\beta}$ if LV (solid line) or PCA (dashed line) is used for microaggregation ($\sigma_\epsilon = 0.5$)

Fig. 2 shows what happens when σ_ϵ is increased: As σ_ϵ increases, the bias increases, and also the difference between the two bias curves. In addition, for all values of σ_ϵ, β is underestimated by $\hat{\beta}$ if $\beta < 0$.

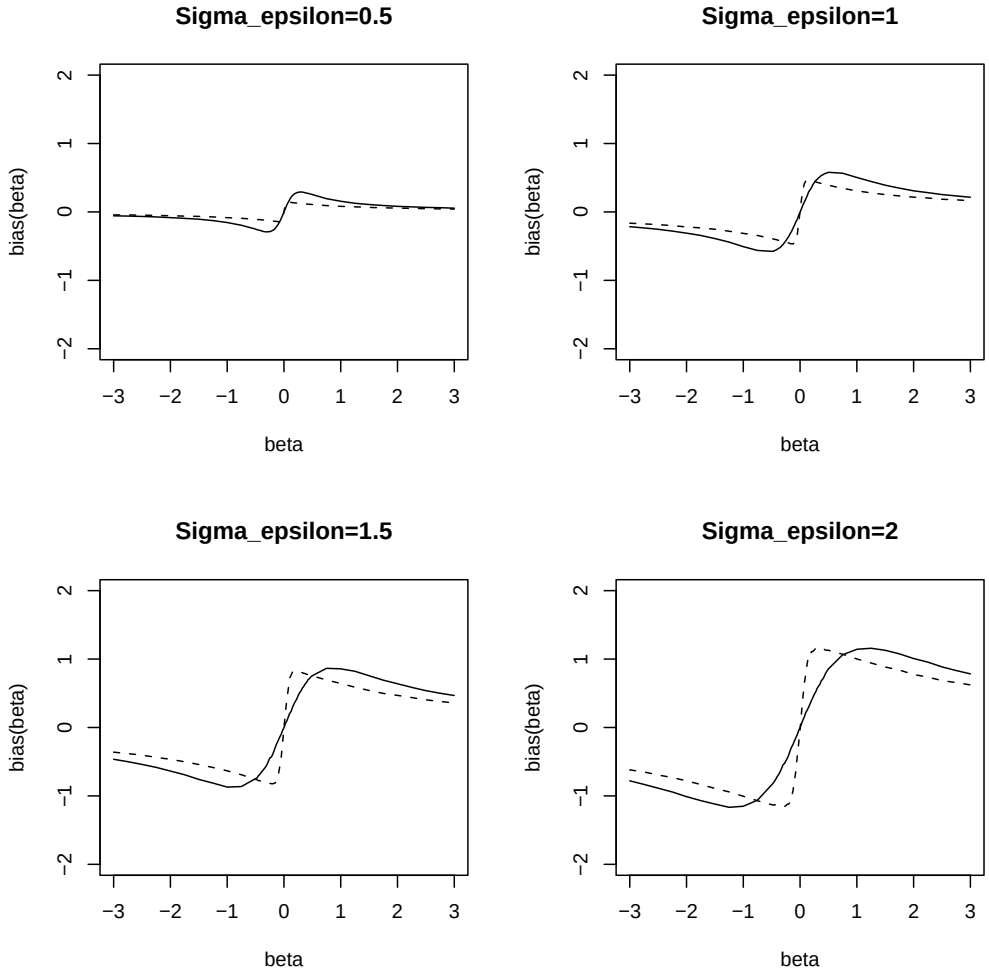

Figure 2: Bias of $\hat{\beta}$ if LV (solid line) or PCA (dashed line) is used for microaggregation

4 Estimating a Linear Model with Two Covariates

In this section we outline the effects of microaggregation on the estimation of a multiple linear regression model. We therefore expand model (1) by adding an additional covariate X_2:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \ . \tag{2}$$

Again, we study the effects of the five microaggregation techniques described in Section 2 by means of a systematic simulation study. The values of the covariates X_1 and X_2 were independently drawn from a standard normal distribution, the number of replications was 1000. The residual variance σ_ϵ^2 was set equal to one. We estimated model (2) for various samples sizes, setting $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = -0.5$, and $A = 3$. Tables 5 and 6 show the results obtained. Apparently the results of Section 3 (concerning the estimation of the simple linear model (1)) can be applied to the estimation of the multiple linear regression model (2) as well. The naive estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ based on

		n=50	n=100	n=200	n=400
$\beta_1 = 1$	Eucl	1.094	1.068	1.046	1.030
	IR	0.989	0.986	1.001	1.001
	LV	1.439	1.443	1.425	1.422
	PCA	1.264	1.258	1.265	1.261
	kW	1.149	1.107	1.063	1.044

Table 5: Naive estimates of β_1 for various sample sizes ($\sigma_\epsilon = 1$)

		n=50	n=100	n=200	n=400
$\beta_2 = -0.5$	Eucl	-0.569	-0.544	-0.523	-0.520
	IR	-0.495	-0.504	-0.500	-0.501
	LV	-0.725	-0.710	-0.720	-0.713
	PCA	-0.617	-0.636	-0.624	-0.629
	kW	-0.592	-0.547	-0.530	-0.520

Table 6: Naive estimates of β_2 for various sample sizes ($\sigma_\epsilon = 1$)

Eucl, IR, and kW are biased if the sample size n is small. However, Eucl, IR, and kW lead to consistent estimates of the slope parameters β_1 and β_2. In contrast, $\hat{\beta}_1$ and $\hat{\beta}_2$ are asymptotically biased if LV and PCA are used to aggregate the data.

To explore how the asymptotic bias induced by LV and PCA depends on the regression parameter values, we estimated β_1 and β_2 for $\beta_1 = 0, 0.005, \ldots, 0.495, 0.5, 0.75, \ldots, 2.75, 3$, $\beta_2 = 0, 0.005, \ldots, 0.495, 0.5, 0.75, \ldots, 2.75, 3$, and the corresponding negative values. As in Section 3, the sample size n was set equal to 400 and the group size A was set equal to three. The values of X_1 and X_2 were drawn independently from a standard normal distribution. The results are shown in Figs. 3 (LV) and 4 (PCA): (Note that for symmetry reasons, bias($\hat{\beta}_2$) is exactly the same as bias($\hat{\beta}_1$). This is why we do not include the graphs showing bias($\hat{\beta}_2$)). We see that, just as in Figs. 1 and 2, bias($\hat{\beta}_1$) is positive if β_1 is positive and bias($\hat{\beta}_1$) is negative if β_1 is negative. Moreover, if one of the parameters β_1 and β_2 is "large" in absolute value, bias($\hat{\beta}_1$) and bias($\hat{\beta}_2$) both go to zero. This result can be explained by the fact that if either β_1 or β_2 is large in absolute value, LV and PCA are almost the same as microaggregation using X_1 or X_2 as the leading variable. In this case, the naive estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased (see Feige/Watts 1972).

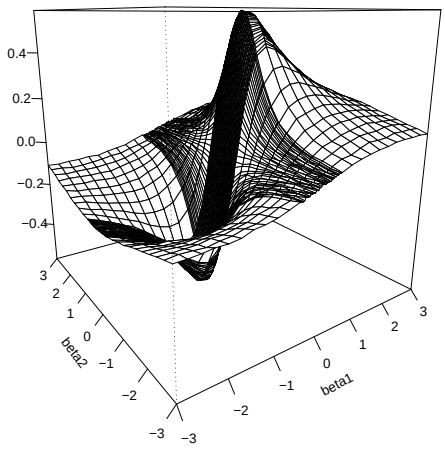

Figure 3: Plot of bias($\hat{\beta}_1$) vs. β_1 and β_2 if LV is used for microaggregation

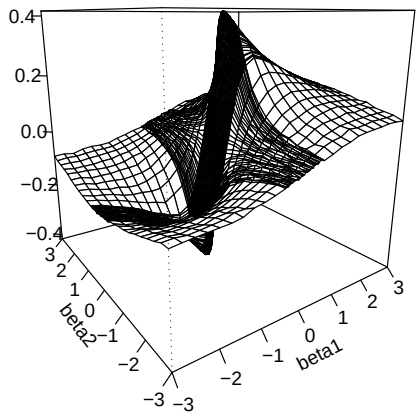

Figure 4: Plot of bias($\hat{\beta}_1$) vs. β_1 and β_2 if PCA is used for microaggregation

5 Conclusion

We have analyzed the effects of microaggregation techniques on the estimation of a linear model, one of the most frequently encountered statistical estimation problems. By means of a simulation study, we have investigated the behavior of the naive least-squares estimator of the slope parameter in a simple linear model based on microaggregated data.

The main results are:

1. All five microaggregation techniques considered in this paper induce a bias in the naive estimator of the slope parameter β in model (1). If IR is used to aggregate the data and $\beta > 0$, the bias is negative, whereas if LV, PCA, Eucl, or kW are used for aggregation and $\beta > 0$, the bias is positive. For symmetry reasons, if $\beta < 0$, the bias is positive for IR and negative for LV, PCA, Eucl, and kW. If $\beta = 0$, $\hat{\beta}$ is unbiased despite microaggregation.

2. Concerning the asymptotic behavior of the naive least-squares estimator, the simulation results show that for IR, Eucl, and kW the bias disappears for large sample sizes. Thus, IR, Eucl, and kW lead to consistent estimates of β.

3. If LV and PCA are used to mask the data, estimates are asymptotically biased. The asymptotic bias is a non-monotonic function of β and converges to 0 as $|\beta| \to \infty$. For small values of $|\beta|$, the estimates based on PCA show a larger bias than those based on LV. If $|\beta|$ is large, PCA induces a smaller bias than LV does.

4. The bias of the estimates becomes larger if the residual variance σ_ϵ^2 is increased. In addition, if IR, Eucl, or kW are used for microaggregation, convergence to the true value of β slows down. The same effect can be seen when the group size A is increased (except for the estimates based on Eucl, which do not seem to depend on the group size).

5. The above results are also found in a multiple linear regression model with two covariates. In addition, if any one of the two slope parameters β_1 or β_2 is large in absolute value, the bias of the naive estimates is close to zero.

We thus see that in terms of consistency, Eucl and kW are reasonable alternatives to IR which is considered to be less protective than the other microaggregation techniques analyzed in this paper (see Winkler 2002). However, Eucl and kW clearly are the most computationally demanding microaggregation procedures. Concerning LV and PCI, we have seen that estimates are asymptotically biased. An obvious solution to this problem would be to develop estimators that correct for the biases induced by LV and PCI. This can be done if the bias has been evaluated analytically. For LV this has been achieved, see Schmid et al. (2005).

Acknowledgements

We gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (German Science Foundation).

References

Defays, D., M.N. Anwar (1998), Masking Microdata Using Micro-Aggregation. Journal of Official Statistics, Vol. 14, No. 4, pp. 449-461.

Domingo-Ferrer, J., J.M. Mateo-Sanz (2001), An Empirical Comparison of SDC Methods for Continuous Microdata in Terms of Information Loss and Re-Identification Risk. Second Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, Macedonia, March 14-16, 2001.

Domingo-Ferrer, J., J.M. Mateo-Sanz (2002), Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 1, pp. 189-201.

Domingo-Ferrer, J., V. Torra (2001), A Quantitative Comparison of Disclosure Control Methods for Microdata. In: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (eds.), Confidentiality, Disclosure, and Data Access, North-Holland, Amsterdam, pp. 111-133.

Elliot, M. (2001), Disclosure Risk Assessment. In: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (eds.), Confidentiality, Disclosure, and Data Access, North-Holland, Amsterdam, pp. 75-90.

Feige, E.L., H.W. Watts (1972), An Investigation of the Consequences of Partial Aggregation of Micro-Economic Data. Econometrica, Vol. 40, No. 2, pp. 343-360.

Mateo-Sanz, J.M., J. Domingo-Ferrer (1998), A Comparative Study of Microaggregation Methods. Questiio, Vol. 22, No. 3, pp. 511-526.

Paass, G., U. Wauschkuhn (1985), Datenzugang, Datenschutz und Anonymisierung - Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten. Berichte der Gesellschaft für Mathematik und Datenverarbeitung, Nr. 148, Oldenbourg, München.

Schmid, M., H. Schneeweiss, H. Küchenhoff (2005), Consistent Estimation of a Simple Linear Model Under Microaggregation. Discussion Paper 415, SFB 386, Institut für Statistik, Ludwig-Maximilians-Universität München.

Willenborg, L., T. de Waal (2001), Elements of Statistical Disclosure Control. Springer, New York.

Winkler, W.E. (2002), Single-Ranking Micro-aggregation and Re-identification. Statistical Research Division report RR 2002/08, U.S. Bureau of the Census, Washington.

Yancey, W.E., W.E. Winkler, R.H. Creecy (2002), Disclosure Risk Assessment in Perturbative Microdata Protection. In J. Domingo-Ferrer (ed.), Inference Control in Statistical Databases, Springer, New York, pp. 135-152.