Journal of Official Statistics, Vol. 25, No. 4, 2009, pp. 529–548

The Effect of Single-Axis Sorting on the Estimation of a Linear Regression

Matthias Schmid[1]

Microaggregation is one of the most important statistical disclosure control techniques for continuous microdata. Observations in a data set are grouped and replaced by their corresponding group means, so that identification of sensitive observations is unlikely. However, microaggregation is also known to affect the results of statistical analyses. In this article we investigate the impact of microaggregation on the least squares estimation of a linear model in continuous variables. It is shown that least squares estimators are not necessarily consistent if the groups of observations are formed by means of a sorting variable. Using this result, we develop a consistent estimator that removes the aggregation bias. Moreover, we derive the asymptotic covariance matrix of the corrected least squares estimator.

Key words: Asymptotic variance; consistent estimation; disclosure control; linear model; microaggregation; sorting variable.

1. Introduction

A problem statistical offices are increasingly faced with is guaranteeing the confidentiality of respondents when releasing microdata sets. This problem is usually solved by anonymizing confidential data with the help of disclosure control techniques. One of the most important disclosure control techniques for continuous data is microaggregation, a method where observations are grouped and replaced by their respective group means (Defays and Nanopoulos 1993; Defays and Anwar 1998; Domingo-Ferrer and Mateo-Sanz 2002). Each group is forced to contain a minimum number of at least K observations. Thus, as each observation in the microaggregated data set appears at least K times, observations cannot be identified, and the disclosure risk of the anonymized data is kept low. This concept, which is commonly referred to as "K-anonymity" (Samarati 2001; Sweeney 2002), is one of the key concepts of modern data protection and has recently been extended by Machanavajjhala et al. (2006) and Li et al. (2007).

In the literature, various suggestions have been made as to how to form the groups for microaggregation (Defays and Nanopoulos 1993; Defays and Anwar 1998; Mateo-Sanz and Domingo-Ferrer 1998; Domingo-Ferrer and Mateo-Sanz 2002; Laszlo and Mukherjee 2005; Domingo-Ferrer et al. 2006; Solanas et al. 2006; Solanas and Martinez-Balleste 2006; Domingo-Ferrer et al. 2008). We consider one of the most popular microaggregation techqniques, namely microaggregation by single-axis sorting (Mateo-Sanz and Domingo-Ferrer 1998). With this technique, the observations in a data set are first sorted by a

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-University Erlangen-Nuremberg, Waldstrasse 6, 91054 Erlangen, Germany. Email: matthias.schmid@imbe.med.uni-erlangen.de

prespecified sorting variable (such as the first principal component projection of the data or the sum of z-scores). The sorted data set is then subdivided into small groups consisting of K consecutive observations each.

While microaggregation has turned out to be an effective tool for protecting sensitive observations in a data set (Ronning et al. 2005; Lenz 2006, Section 8), the technique also affects the results of statistical analyses. This is particularly true for microaggregation by single-axis sorting: Empirical analyses and simulation studies have shown that parameter estimates might be biased and/or less efficient if computed from the microaggregated data (Schmid and Schneeweiss 2005). Investigating the impact of microaggregation on the results of statistical analyses is thus an important task, as the release of microaggregated data sets only makes sense if statistical analyses based on these data sets are analytically valid.

In this article we carry out a theoretical analysis of the effect of single-axis sorting on the least squares (LS) estimation of a linear model in continuous variables. The starting point of the article is the work of Schmid et al. (2007), who considered the case where the dependent variable in the linear model is used as the sorting variable. Schmid et al. have shown that in this case the LS estimator of a linear model is asymptotically biased but can be corrected such that the estimation procedure becomes consistent. In the following, we will generalize the results of Schmid et al. (2007) to the case where an *arbitrary* sorting variable H is used for microaggregation. Considering the case of an arbitrary sorting variable H instead of the dependent variable of the linear model is important because data holders often use sorting variables which are linear combinations of the variables contained in a data set (such as the first principal component projection) or even sorting variables which are not contained in the linear model at all. As a consequence, the results derived in this article can be more widely applied than the results presented in Schmid et al. (2007).

We will first derive the asymptotic properties of the standard LS estimators when applied to a data set that has been microaggregated with respect to H. After having shown that the LS estimators are asymptotically biased, we will develop a new estimation procedure that corrects for the bias, leading to a consistent estimator of the linear model. The asymptotic covariance matrix of the corrected LS estimator of the slope parameter vector β will also be derived.

Section 2 starts with a brief description of microaggregation by single-axis sorting. In Section 3 we derive theoretical results on the effect of this technique on the LS estimation of a linear model. Furthermore, a method for correcting the aggregation bias is developed. Section 4 deals with the asymptotic covariance matrix of the corrected LS estimator. Section 5 contains a simulation study on the results derived in Sections 3 and 4. A summary of the article is given in Section 6.

2. Microaggregation by Single-axis Sorting

An obvious question arising from the concept of K-anonymity is how to form the groups that are used for microaggregation. Clearly, disseminating the group means instead of the original data values will almost always result in a loss of information. In the context of microaggregation it is common to use the trace of the empirical within-group covariance matrix as a measure of information loss (see Domingo-Ferrer and Mateo-Sanz 2002). This implies that the optimal partition of the data consists of "homogeneous" groups of at least K observations each, where the data values in each group deviate as little as possible from the corresponding group means.

Since finding the optimal partition has proved to be NP-hard if there are more than two variables in a data set (Oganian and Domingo-Ferrer 2001), a variety of heuristic microaggregation techniques have been proposed in the literature. Some of these techniques, such as nonhierarchical K-partitioning (Defays and Nanopoulos 1993) and microaggregation based on Euclidean distances (Domingo-Ferrer and Mateo-Sanz 2002), work with a fixed group size K. On the other hand there are a number of microaggregation techniques that result in variable-sized groups with minimum cardinality K. These techniques, which are termed "data-oriented microaggregation techniques," include K-Ward microaggregation (Domingo-Ferrer and Mateo-Sanz 2002), MST partitioning (Laszlo and Mukherjee 2005), microaggregation based on genetic algorithms (Solanas et al. 2006), multivariate extensions of the Hansen-Mukherjee method for optimal one-dimensional microaggregation (Hansen and Mukherjee 2003; Domingo-Ferrer et al. 2006), variable-size MDAV (Solanas and Martinez-Balleste 2006), and the μ-Approximation algorithm (Domingo-Ferrer et al. 2008). Microaggregation techniques with variable-sized groups usually result in a close approximation of the optimal partition of a data set. However, their computational cost is typically higher than the cost of fixed-size microaggregation techniques.

In this article we consider microaggregation by single-axis sorting (Mateo-Sanz and Domingo-Ferrer 1998), which is one of the most popular and computationally most efficient microaggregation techniques. Single-axis sorting microaggregation uses a fixed group size K, as well as a prespecified sorting variable for partitioning the observations in a data set. The technique works as follows: First, the data set is sorted with respect to the sorting variable. After a fixed group size K has been chosen, the sorted data set is subdivided into small groups consisting of K consecutive observations each. If the sample size n is not a multiple of K, it is a common strategy to assign $K + \text{mod}(n/K)$ observations to the group around the median of the sorting variable. In the following, we assume that the sample size n is a multiple of K (the asymptotic results presented in Sections 3 and 4 are not affected by a slightly larger "median" group containing $K + \text{mod}(n/K)$ observations). After the groups have been formed, the data in each of the n/K groups are averaged, and the averages are assigned to the observations of the respective groups. In practice, K is usually chosen to be 3 or 5. The sorting variable should be chosen such that it reflects the multivariate structure of the data set. This requirement ensures that groups of consecutive observations are homogeneous, so that the loss of information due to averaging the data is kept small.

Microaggregation is most often used as a disclosure control technique for continuous data, so we exclusively consider continuous data sets in this article (disclosure control techniques for discrete data can be found in e.g., Willenborg and de Waal 2001 and Doyle et al. (2001); microaggregation techniques for discrete data have been developed by Domingo-Ferrer and Torra (2005)).

As an example of single-axis sorting we consider a data set with six observations and three variables X_1, X_2, and Y:

x_1	2.00	1.00	5.00	9.00	3.00	4.00
x_2	1.00	3.00	4.00	2.00	8.00	6.00
y	2.00	7.00	6.00	8.00	3.00	1.00

Assume the sorting variable H to be the first principal component projection based on the correlation matrix of the data and set $K = 3$. Then the data values of H are given by $h = (-0.38, 0.20, 0.54, 2.16, -1.32, -1.20)$. Sorting with respect to H yields the sorted data set

$x_{1,sort}$	3.00	4.00	2.00	1.00	5.00	9.00
$x_{2,sort}$	8.00	6.00	1.00	3.00	4.00	2.00
y_{sort}	3.00	1.00	2.00	7.00	6.00	8.00
(h_{sort})	(-1.32)	(-1.20)	(-0.38)	(0.20)	(0.54)	(2.16)

and the microaggregated data set

$\tilde{x}_1$	3.00	3.00	3.00	5.00	5.00	5.00
$\tilde{x}_2$	5.00	5.00	5.00	3.00	3.00	3.00
$\tilde{y}$	2.00	2.00	2.00	7.00	7.00	7.00

3. Consistent Estimation of the Parameters of a Linear Regression Model

3.1. Notation

We consider the effect of single-axis sorting on the LS estimation of the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \tag{1}$$

Y denotes the continuous response variable with variance σ_{yy}. $X_1, \ldots, X_p$ denote the continuous regressors with variances $\sigma_{11}, \ldots, \sigma_{pp}$, respectively, and ε denotes the random error with zero mean and variance σ_ε^2. β_0 is the intercept and $\beta := (\beta_1, \ldots, \beta_p)'$ is the vector of slope parameters in Model (1). Further consider a continuous random variable H with variance σ_{hh}. We assume $Y, X_1, \ldots, X_p$, and H to be jointly normally distributed. Then ε is also normally distributed and is independent of $(X_1, \ldots, X_p)$. The objective is to estimate β and σ_ε^2 from an i.i.d. sample with n observations $(y_v, x_{v1}, \ldots, x_{vp})$, $v = 1, \ldots, n$, that has been microaggregated with respect to the sorting variable H. As noted before, we assume that n is a multiple of the group size K.

Let $y := (y_1, \ldots, y_n)'$ and $x_i := (x_{1i}, \ldots, x_{ni})'$, $i = 1, \ldots, p$, contain the original data values. Further let $h := (h_1, \ldots, h_n)'$ contain the original data values of H. The vectors containing the aggregated data are denoted by $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$, and $\tilde{h}$. Note that the empirical means $\bar{y}, \bar{x}_1, \ldots, \bar{x}_p, \bar{h}$ of $y, x_1, \ldots, x_p, h$ are the same as the empirical means $\bar{\tilde{y}}, \bar{\tilde{x}}_1, \ldots, \bar{\tilde{x}}_p, \bar{\tilde{h}}$ of $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p, \tilde{h}$, respectively.

Throughout this article we use the following notations: The covariance of two arbitrary random variables U and V is denoted by σ_{uv} and the variance of U by $\sigma_{uu} = \sigma_u^2$. If $W := (W_1, \ldots, W_p)'$ is a vector of random variables, the column covariance vector with elements σ_{w_iu}, $i = 1, \ldots, p$, is denoted by σ_{wu}. The covariance matrix of W is denoted by Σ_{ww}. In case $W = (X_1, \ldots, X_p)'$, we use the abbreviations $\sigma_{ij} := \sigma_{x_ix_j}$ and $\sigma_{iu} := \sigma_{x_iu}$.

The corresponding empirical concepts for the n-tupels $u = (u_1, \ldots, u_n)'$, $v = (v_1, \ldots, v_n)'$, and $w_i = (w_{1i}, \ldots, w_{ni})'$, $i = 1, \ldots, p$, are the empirical covariance of u and v: $s_{uv} := 1/n\sum_{\nu=1}^{n} (u_\nu - \bar{u})(v_\nu - \bar{v})$, the empirical variance of u: $s_u^2 := s_{uu}$, the empirical covariance vector s_{wu} with elements $s_{w_{iu}}$, and the empirical covariance matrix s_{ww} with elements $s_{w_iw_j}$. In case $w_i = (x_{1i}, \ldots, x_{ni})'$, the abbreviations $s_{ij} := s_{x_ix_j}$ and $s_{iu} := s_{x_iu}$ are used.

The corresponding expressions for microaggregated data are denoted by a tilde on top of u and v. For example, the empirical covariance of the n-tupels $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_n)'$ and $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_n)'$ containing the microaggregated values of u and v, respectively, is denoted by $s_{\tilde{u}\tilde{v}}$. We also use the abbreviations $s_{\tilde{x}_i\tilde{x}_j} =: s_{\widetilde{ij}}$ and $s_{\tilde{x}_i\tilde{y}} =: s_{\widetilde{iy}}$. Finally, we denote the probability limits ($n \longrightarrow \infty$) of $s_{\tilde{u}\tilde{v}}$, $s_{\tilde{w}\tilde{u}}$, and $S_{\tilde{w}\tilde{w}}$ by $\sigma_{\tilde{u}\tilde{v}}$, $\sigma_{\tilde{w}\tilde{u}}$, and $\Sigma_{\tilde{w}\tilde{w}}$, respectively.

3.2. *Examples of Sorting Variables*

Due to the joint normality of $Y, X_1, \ldots, X_p$, and H, the sorting variable can be expressed as

$$H = c_yY + c_1X_1 + \cdots + c_pX_p + \varphi \tag{2}$$

where $c := (c_y, c_1, \ldots, c_p)'$ is a vector of coefficients and φ is a normally distributed error variable with zero mean that is independent of $Y, X_1, \ldots, X_p$. (Alternatively, it could have been assumed in Section 3.1 that (2) holds and that $Y, X_1, \ldots, X_p$ are jointly normally distributed. This would imply the joint normality of $Y, X_1, \ldots, X_p$, and H).

In practice, the sorting variable will most often be a linear combination of the variables in Model (1), implying that $\varphi \equiv 0$. Popular choices for the sorting variable include (1) the dependent variable Y (where $\varphi \equiv 0$, $c_y = 1$, $c_1 = \ldots = c_p = 0$), (2) a regressor X_i (where $\varphi \equiv 0$, $c_y = 0$, $c_1 = \ldots = c_{i-1} = 0$, $c_i = 1$, $c_{i+1} = \ldots = c_p = 0$), (3) the first principal component projection of $Y, X_1, \ldots, X_p$ (where $\varphi \equiv 0$ and c is the eigenvector associated with the largest eigenvalue of the covariance or correlation matrix of $Y, X_1, \ldots, X_p$), (4) the sum of z-scores of the variables in the linear model (where $\varphi \equiv 0$, $c_y = \sigma_{yy}^{-1/2}$, $c_1 = \sigma_{11}^{-1/2}, \ldots, c_p = \sigma_{pp}^{-1/2}$), and (5) an arbitrary variable H which is included in the original data set (and possibly also in the released data set) but not in the linear regression model (1) (in this case, typically, $\varphi \neq 0$).

Concerning the effect of single-axis sorting on the LS estimation of Model (1), the following results have been derived in the literature: The LS estimator of β is unbiased if a regressor (or any function that solely depends on the regressors) is used as the sorting variable (Feige and Watts 1972). By contrast, if the dependent variable Y is used as the sorting variable, the LS estimator of β is asymptotically biased (Schmid et al. 2007). Schmid et al. also derived a consistent estimator of β that corrects for the aggregation bias.

In the following sections we generalize these results to the case where an arbitrary sorting variable H is used for single-axis sorting.

3.3. Consistent Estimation of β

We focus on the estimation of the vector of genuine regression coefficients $\beta = (\beta_1, . . ., \beta_p)'$. When we know how to estimate β consistently, it will be clear how to estimate β_0 and σ_ε^2 as well. We denote the least squares estimator of β by $\tilde{b}$, which is given by

$$\tilde{b} := S_{\tilde{x}\tilde{x}}^{-1} s_{\tilde{x}\tilde{y}} \tag{3}$$

In order to study the bias of $\tilde{b}$ and to construct a consistent estimator of β, we need the following lemma:

Lemma 1. Consider the set of inverse linear regressions

$$X_i = \alpha_i + \gamma_i H + \delta_i, \;\; i = 1, . . . , p \tag{4}$$

$$Y = \alpha_y + \gamma_y H + \delta_y \tag{5}$$

with $\mathrm{E}(\delta_i) = 0$, $i = y, 1, . . ., p$. Then the following probability limits exist:

$$a) \;\; \operatorname*{plim}_{n\to\infty} s_{\tilde{h}\tilde{h}} = \sigma_{hh}$$

$$b) \;\; \operatorname*{plim}_{n\to\infty} s_{\tilde{x}\tilde{h}} = \sigma_{xh}$$

$$c) \;\; \operatorname*{plim}_{n\to\infty} s_{\tilde{y}\tilde{h}} = \sigma_{yh}$$

$$d) \;\; \operatorname*{plim}_{n\to\infty} S_{\tilde{x}\tilde{x}} = \Sigma_{\tilde{x}\tilde{x}} = \frac{1}{K}\Sigma_{xx} + \left(1 - \frac{1}{K}\right)\frac{\sigma_{xh}\sigma'_{xh}}{\sigma_{hh}}$$

$$e) \;\; \operatorname*{plim}_{n\to\infty} s_{\tilde{x}\tilde{y}} = \sigma_{\tilde{x}\tilde{y}} = \frac{1}{K}\sigma_{xy} + \left(1 - \frac{1}{K}\right)\frac{\sigma_{yh}}{\sigma_{hh}}\sigma_{xh}$$

Proof. As H is a sorting variable related to $Y, X_1, . . . , X_p$ by the linear regression Models (4) and (5), Lemma 1 follows directly from Lemma 1 in Schmid et al. (2007). The basic idea of the proof is to decompose the variances/covariances computed from the original data into the variances/covariances between the groups (which are equal to the variances/covariances computed from the aggregated data) and the variances/covariances within the groups. Relations a) to c) can then be derived by showing that the latter expressions converge to zero in probability as $n \to \infty$. The probability limits in d) and e) can be derived with the help of the inverse regressions (4) and (5), which exist because of the joint normality of $Y, X_1, . . . , X_p$, and H. The key role of (4) and (5) in the proof of d) and e) is understandable when one remembers that microaggregation with respect to a regressor – and in (4) and (5), H is the regressor variable – does not result in any bias of the LS estimator (Feige and Watts 1972). □

With Lemma 1, the probability limit of $\tilde{b}$ and hence its asymptotic bias can be determined:

Theorem 1. Under the assumptions of Section 3.1, the probability limit of the LS estimator $\tilde{b}$ of the parameter vector β in Model (1) is given by

$$
\begin{aligned}
\tilde{\beta} &:= \operatorname*{p\,lim}_{n\to\infty} \tilde{b} \\
&= \left(\Sigma_{xx}^{-1} - \Sigma_{xx}^{-1} \frac{(K-1)/\sigma_{hh}\sigma_{xh}\sigma_{xh}^{'}}{1+(K-1)/\sigma_{hh}\sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xh}} \Sigma_{xx}^{-1} \right) \left(\sigma_{xy} + \frac{K-1}{\sigma_{hh}} \sigma_{yh}\sigma_{xh} \right)
\end{aligned}
\tag{6}
$$

Proof. Define $a := (K-1)/\sigma_{hh}$. By Lemma 1 and from the definition of $\tilde{b}$ in (3), it follows that

$$
\begin{aligned}
\tilde{\beta} = \Sigma_{\tilde{x}\tilde{x}}^{-1}\sigma_{\tilde{x}\tilde{y}} &= (\Sigma_{xx} + a\sigma_{xh}\sigma_{xh}^{\prime})^{-1}(\sigma_{xy} + a\sigma_{yh}\sigma_{xh}) \\
&= \left(\Sigma_{xx}^{-1} - \Sigma_{xx}^{-1} \frac{a\sigma_{xh}\sigma_{xh}^{'}}{1+a\sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xh}} \Sigma_{xx}^{-1} \right) (\sigma_{xy} + a\sigma_{yh}\sigma_{xh})
\end{aligned}
\tag{7}
$$

which proves the theorem. Note that in order to obtain (7), we used the Sherman-Morrison matrix inversion formula (cf. Dhrymes 1984, Corollary 5). □

Using $\beta = \Sigma_{xx}^{-1}\sigma_{xy}$, it follows from (7) that

$$
\begin{aligned}
\tilde{\beta} &= \beta + a\sigma_{yh}\Sigma_{xx}^{-1}\sigma_{xh} - \frac{a\sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xy}}{1+a\sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xh}}\Sigma_{xx}^{-1}\sigma_{xh} \quad - \frac{a^{2}\sigma_{yh}\sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xh}}{1+a\sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xh}}\Sigma_{xx}^{-1}\sigma_{xh} \\
&= \beta + \frac{a(\sigma_{yh} - \sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xy})}{1+a\sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xh}}\Sigma_{xx}^{-1}\sigma_{xh}
\end{aligned}
\tag{8}
$$

From (8) we obtain the important result that the LS estimator $\tilde{b}$ is not necessarily a consistent estimator of β. As expected, in case of the nonaggregated data (i.e., $K = 1$), the asymptotic bias of $\tilde{b}$ is equal to 0. In addition, if H is uncorrelated with ε (which is the case if one of the regressors or a function of the regressors is the sorting variable), we have $\sigma_{yh} = \beta'\sigma_{xh}$. Thus, in this case, the numerator in (8) becomes

$$
a\Big(\sigma_{yh} - \sigma_{xh}^{'}\Sigma_{xx}^{-1}\sigma_{xy}\Big) = a\Big(\beta'\sigma_{xh} - \sigma_{xh}^{'}\beta\Big) = 0 \tag{9}
$$

implying that the LS estimator $\tilde{b}$ is a consistent estimator of β. This result confirms the work of Feige and Watts (1972), who showed that $\tilde{b}$ is an *unbiased* estimator if H is equal to one of the regressors (or a function of the regressors). Similarly, the bias term in (8) disappears if $\sigma_{xh} = 0$, i.e., if the sorting variable H is independent of the regressors $X_1, \ldots, X_p$.

Finally, from (8), a consistent estimator of β based on the aggregated data can be constructed:

Theorem 2. Under the conditions of Theorem 1, a consistent estimator of β based on the microaggregated data is given by

$$\tilde{b}_c := \tilde{b} + \frac{(K-1)\left(s'_{\tilde{x}\tilde{h}} S^{-1}_{\tilde{x}\tilde{x}} s_{\tilde{x}\tilde{y}} - s_{\tilde{y}\tilde{h}}\right)}{K s_{\tilde{h}\tilde{h}} - (K-1) s'_{\tilde{x}\tilde{h}} S^{-1}_{\tilde{x}\tilde{x}} s_{\tilde{x}\tilde{h}}} S^{-1}_{\tilde{x}\tilde{x}} s_{\tilde{x}\tilde{h}} \tag{10}$$

Proof. We start from $\beta = \Sigma^{-1}_{xx}\sigma_{xy}$ and replace Σ_{xx} by

$$\Sigma_{xx} = \left(K\Sigma_{\tilde{x}\tilde{x}} - (K-1)\frac{\sigma_{xh}\sigma'_{xh}}{\sigma_{hh}}\right) \tag{11}$$

from Lemma 1d). In addition, we replace σ_{xy} by

$$\sigma_{xy} = \left(K\sigma_{\tilde{x}\tilde{y}} - (K-1)\frac{\sigma_{yh}}{\sigma_{hh}}\sigma_{xh}\right) \tag{12}$$

from Lemma 1e). This yields

$$\begin{aligned}
\beta &= \left(K\Sigma_{\tilde{x}\tilde{x}} - (K-1)\frac{\sigma_{xh}\sigma'_{xh}}{\sigma_{hh}}\right)^{-1}\left(K\sigma_{\tilde{x}\tilde{y}} - (K-1)\frac{\sigma_{yh}}{\sigma_{hh}}\sigma_{xh}\right) \\
&= \left(\frac{1}{K}\Sigma^{-1}_{\tilde{x}\tilde{x}} + a\frac{\frac{1}{K}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh}\sigma'_{xh}\frac{1}{K}\Sigma^{-1}_{\tilde{x}\tilde{x}}}{1 - \frac{a}{K}\sigma'_{xh}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh}}\right)(K\sigma_{\tilde{x}\tilde{y}} - a\sigma_{yh}\sigma_{xh}) \\
&= \tilde{\beta} - \frac{a}{K}\sigma_{yh}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh} + \frac{a}{K}\frac{\sigma'_{xh}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{\tilde{x}\tilde{y}}}{1 - \frac{a}{K}\sigma'_{xh}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh}}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh} \\
&\quad - \frac{a^2}{K^2}\frac{\sigma_{yh}\sigma'_{xh}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh}}{1 - \frac{a}{K}\sigma'_{xh}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh}}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh} \\
&= \tilde{\beta} + \frac{(K-1)\left(\sigma'_{xh}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{\tilde{x}\tilde{y}} - \sigma_{yh}\right)}{K\sigma_{hh} - (K-1)\sigma'_{xh}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh}}\Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{xh}
\end{aligned} \tag{13}$$

where $\tilde{\beta} = \Sigma^{-1}_{\tilde{x}\tilde{x}}\sigma_{\tilde{x}\tilde{y}}$ was used. According to Lemma 1, σ_{hh}, σ_{xh}, and σ_{yh} can be consistently estimated by $\sigma_{\tilde{h}\tilde{h}}$, $\sigma_{\tilde{x}\tilde{h}}$, and $\sigma_{\tilde{y}\tilde{h}}$, respectively. A consistent estimator $\tilde{b}_c$ is thus given by

$$\tilde{b}_c = \tilde{b} + \frac{(K-1)\left(s'_{\tilde{x}\tilde{h}} S^{-1}_{\tilde{x}\tilde{x}} s_{\tilde{x}\tilde{y}} - s_{\tilde{y}\tilde{h}}\right)}{K s_{\tilde{h}\tilde{h}} - (K-1) s'_{\tilde{x}\tilde{h}} S^{-1}_{\tilde{x}\tilde{x}} s_{\tilde{x}\tilde{h}}} S^{-1}_{\tilde{x}\tilde{x}} s_{\tilde{x}\tilde{h}} \tag{14}$$

□

Note that the computation of (10) requires the aggregated data values of the sorting variable H to be known to the data users. This either implies that data holders provide the aggregated data values of H or that $\tilde{h}$ can be reconstructed from the aggregated data values

$\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$. Reconstruction of $\tilde{h}$ is, for example, possible for the sorting variables (1) to (4) presented in Section 3.2: Since these sorting variables are exact linear combinations of $X_1, \ldots, X_p$, and Y, the data values of H can be reconstructed from $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$ and from the coefficients $c_y, c_1, \ldots, c_p$:

$$\tilde{h} = c_y\tilde{y} + c_1\tilde{x}_1 + \cdots + c_p\tilde{x}_p \tag{15}$$

It can be seen from (15) that data users only have to know (or estimate) the values of the coefficients $c_y, c_1, \ldots, c_p$ instead of the full vector $\tilde{h}$. Consequently, if one of the variables in the linear model is the sorting variable (i.e., sorting variables (1) or (2) from Section 3.2 are used), it is sufficient to tell data users which variable has been used to sort the data. As in this case all coefficients are equal to 0 except one coefficient (which is associated with the sorting variable and is equal to 1), $\tilde{h}$ is equal to the vector of aggregated values of the variable in the linear model that was used to sort the data. If the first principal component projection of $Y, X_1, \ldots, X_p$ or the sum of z-scores is the sorting variable (i.e., sorting variables (3) or (4) from Section 3.2 are used), the coefficient vector c can be consistently estimated from the microaggregated data by solving a system of nonlinear equations depending on the first and second empirical moments of $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$. For details on how to reconstruct $\tilde{h}$, see Schmid (2007, Section 4.3.3).

A consistent estimator of the intercept β_0 based on the aggregated data is given by

$$\tilde{b}_{0c} := \bar{\tilde{y}} - (\tilde{b}_{1c}\bar{\tilde{x}}_1 + \cdots + \tilde{b}_{pc}\bar{\tilde{x}}_p) \tag{16}$$

where $\tilde{b}_{1c}, \ldots, \tilde{b}_{pc}$ are the elements of $\tilde{b}_c$.

Furthermore, from (11) and from a corresponding formula for σ_{yy}, we obtain a consistent estimator of the residual variance $\sigma_\varepsilon^2 = \sigma_{yy} - \beta'\Sigma_{xx}\beta$ based on the aggregated data:

$$\tilde{s}^2_{\varepsilon,c} := \left(Ks_{\tilde{y}\tilde{y}} - (K-1)\frac{s^2_{\tilde{y}\tilde{h}}}{s_{\tilde{h}\tilde{h}}}\right) - \tilde{b}_c^{\,'}\left(KS_{\tilde{x}\tilde{x}} - (K-1)\frac{s_{\tilde{x}\tilde{h}}s'_{\tilde{x}\tilde{h}}}{s_{\tilde{h}\tilde{h}}}\right)\tilde{b}_c \tag{17}$$

4. Asymptotic Covariance Matrix of $\tilde{b}_c$

In this section we derive the asymptotic covariance matrix of the corrected estimator $\tilde{b}_c$. The following conventions are used: Two random vector sequences a_n and b_n are said to be "asymptotically equivalent," written $a_n \sim b_n$, if $\operatorname*{plim}_{n\to\infty}\sqrt{n}(a_n - b_n) = 0$. The asymptotic covariance matrix of a random vector sequence a_n is said to be "equal to Σ_{aa}/n" if $\operatorname*{plim}_{n\to\infty} a_n =: a_\infty$ exists and if $\sqrt{n}(a_n - a_\infty)$ converges in distribution to $N(0,\Sigma_{aa})$ as $n \longrightarrow \infty$.

First note that by (3) and (10)

$$\tilde{b}_c = F(\tilde{s}) \tag{18}$$

where F is a continuously differentiable function of

$$\tilde{s} := \begin{pmatrix} \text{vech}(S_{\tilde{x}\tilde{x}}) \\ s_{\tilde{x}\tilde{y}} \\ s_{\tilde{x}\tilde{h}} \\ s_{\tilde{y}\tilde{h}} \\ s_{\tilde{h}\tilde{h}} \end{pmatrix} \tag{19}$$

The vector vech($\mathrm{S}_{\tilde{x}\tilde{x}}$) contains the lower triangular elements of the symmetric matrix $S_{\tilde{x}\tilde{x}}$. Denote the probability limit of $\tilde{s}$, which is known from Lemma 1, by $\tilde{\sigma}$. Then

$$\tilde{\sigma} = \begin{pmatrix} \text{vech}(\Sigma_{\tilde{x}\tilde{x}}) \\ \sigma_{\tilde{x}\tilde{y}} \\ \sigma_{xh} \\ \sigma_{yh} \\ \sigma_{hh} \end{pmatrix} \tag{20}$$

The idea is to show that

$$\tilde{s} - \tilde{\sigma} \sim G(s) + \Delta \tag{21}$$

where G is a continuously differentiable function of the second-order moments

$$s := \begin{pmatrix} \text{vech}(S_{xx}) \\ s_{xy} \\ s_{xh} \\ s_{yh} \\ s_{hh} \end{pmatrix} \tag{22}$$

based on the nonaggregated data. The probability limit of $\tilde{s}$ is

$$\sigma := \operatorname*{plim}_{n \to \infty} s = \begin{pmatrix} \text{vech}(\Sigma_{xx}) \\ \sigma_{xy} \\ \sigma_{xh} \\ \sigma_{yh} \\ \sigma_{hh} \end{pmatrix} \tag{23}$$

As will be shown, the "residual vector" Δ is a function of the δ_i's defined in (4) and (5). Moreover, as will also be shown, Δ is asymptotically independent of s. Thus, by computing the covariance matrices of s and Δ and by using the delta method, the asymptotic covariance matrix of $\tilde{s}$ can be derived from (21). From (18), by using the delta method once more, we can finally obtain the asymptotic covariance matrix of $\tilde{b}_c$.

To prove (21), we introduce the following fundamental lemma:

Lemma 2. Consider the inverse linear regression Models (4) and (5), which exist by the joint normality of $Y, X_1, \ldots, X_p$, and H. Denote the empirical variances and covariances of the non-aggregated and aggregated data values of δ_i and δ_j, *i, j,* = *y, 1, . . ., p,* by $s_{\delta_i \delta_j}$ and $s_{\tilde{\delta}_i \tilde{\delta}_j}$, respectively. Then the following relations hold for *i, j,* = *y, 1, . . ., p:*

a) $s_{\tilde{i}\tilde{h}} - \sigma_{ih} \sim s_{ih} - \sigma_{ih}$

b) $s_{\tilde{h}\tilde{h}} - \sigma_{hh} \sim s_{hh} - \sigma_{hh}$

c) $s_{\tilde{i}\tilde{j}} - \sigma_{\tilde{i}\tilde{j}} \sim \dfrac{1}{K}(s_{ij} - \sigma_{ij}) + \left(1 - \dfrac{1}{K}\right)\left(\dfrac{s_{ih}s_{jh}}{s_{hh}} - \dfrac{\sigma_{ih}\sigma_{jh}}{\sigma_{hh}}\right) + \left(s_{\tilde{\delta}_i \tilde{\delta}_j} - \dfrac{1}{K} s_{\delta_i \delta_j}\right)$

Proof. As H is a normally distributed random variable related to $Y, X_1, \ldots, X_p$ by the inverse linear regressions (4) and (5), Lemma 2 is a direct consequence of Lemma 2 in Schmid et al. (2007). Again, the variances/covariances computed from the original data can be decomposed into the variances/covariances between the groups (which are equal to the variances/covariances computed from the aggregated data) and the variances/covariances within the groups. Relations a) to c) can then be derived by multiplying the variance/covariance components with $\sqrt{n}$ and by analyzing the probability limits of the differences between the left-hand sides and the right-hand sides of a) to c). For a detailed proof we refer to Schmid et al. (2007) and Schmid (2007). □

Lemma 2 can be used to define the elements of Δ: Let

$$S_{\delta,xx} := \left(s_{\tilde{\delta}_i \tilde{\delta}_j} - \frac{1}{K} s_{\delta_i \delta_j}\right)_{i,j=1,\ldots,p} \tag{24}$$

$$s_{\delta,xy} := \left(s_{\tilde{\delta}_i \tilde{\delta}_y} - \frac{1}{K} s_{\delta_i \delta_y}\right)_{i=1,\ldots,p} \tag{25}$$

Then

$$\Delta := \begin{pmatrix} \mathrm{vech}(S_{\delta,xx}) \\ s_{\delta,xy} \\ \mathbf{0} \end{pmatrix} \tag{26}$$

where $\mathbf{0}$ is a $(p+2)$-dimensional vector of zeros. From Lemma 2 and from the definition of the elements of Δ, it is now clear that Equation (21) holds: The function G is implicitly given by the right-hand sides of the relations a), b), and c) of Lemma 2, but without the term $s_{\tilde{\delta}_i \tilde{\delta}_j} - 1/K s_{\delta_i \delta_j}$. We next show that $G(s)$ and Δ are asymptotically independent and that the asymptotic covariance matrix of Δ can be evaluated:

Lemma 3. Under the conditions of Lemma 2:
a) Let s and σ be as in (22) and (23). Then the vector $\sqrt{n}(\Delta', (s - \sigma)')$ is asymptotically normally distributed. Moreover, the vectors Δ and s are asymptotically independent.

b) Let $\Delta_{ij} := (s_{\tilde{\delta}_i\tilde{\delta}_j} - s_{\delta_i\delta_j}/K)$, $i, j = y, 1, \ldots, p$. Then the asymptotic covariance of Δ_{ij} and Δ_{lm}, $i, j, l, m = y, 1, \ldots, p$, is given by

$$\sigma_{\Delta_{ij}\Delta_{lm}} := \frac{1}{n}\frac{K-1}{K^2}(\sigma_{\delta_i\delta_l}\sigma_{\delta_j\delta_m} + \sigma_{\delta_i\delta_m}\sigma_{\delta_j\delta_l}) \tag{27}$$

Proof. Lemma 3 follows directly from Lemma 3 in Schmid et al. (2007), where H was set equal to the (normally distributed) dependent variable Y. As Y is related to $X_1, \ldots, X_p$ in the same way as H is related to $Y, X_1, \ldots, X_p$ (see (2)), the same proof techniques as in Schmid et al. (2007) can be applied. For details on the proof we refer to Schmid et al. (2007) and Schmid (2007). □

With the help of Lemma 3, the covariance matrix of Δ (denoted by Σ_Δ) can be evaluated. Note that the elements of Σ_Δ corresponding to the zero subvector of Δ are equal to 0.

By applying the delta method, we obtain from (21) and from Lemma 3 that

$$\mathrm{cov}\,(\tilde{s}) = D_G\,\mathrm{cov}\,(s)D_G^{'} + \Sigma_\Delta \tag{28}$$

where D_G is the Jacobian of $G(s)$ evaluated at $\operatorname*{plim}_{n\to\infty} s = \sigma$.

The covariance matrix of s in (28) can be derived as follows: Denote the covariance matrix of $(Y, X_1, \ldots, X_p, H)$ by $\Sigma_{Y,X,H}$ and the empirical covariance matrix of $(Y, X_1, \ldots, X_p, H)$ by $S_{Y,X,H}$. Now, as $n{\cdot}S_{Y,X,H}$ follows a Wishart $(p+2, n-1, \Sigma_{Y,X,H})$ distribution, we have

$$\mathrm{cov}\,(s_{ij}, s_{lm}) = \frac{1}{n}(\sigma_{il}\sigma_{jm} + \sigma_{im}\sigma_{jl}),\ i,j,l,m = y, 1, \ldots, p, h \tag{29}$$

(cf. Evans et al. 1993, p. 158).

From (18) and (28), by applying the delta method once more, we finally obtain

Theorem 3. Under the conditions of Lemma 2, the asymptotic covariance matrix of the corrected estimator $\tilde{b}_c$ is given by

$$\mathrm{cov}\,(\tilde{b}_c) = D_F(D_G\,\mathrm{cov}\,(s)D_G' + \Sigma_\Delta)D_F' \tag{30}$$

where D_F is the Jacobian of $F(\tilde{s})$ evaluated at $\tilde{\sigma}$. $\mathrm{cov}\,(\tilde{b}_c)$ can be estimated by replacing

- σ_{ih}, $i = y, 1, \ldots, p$, with their consistent estimators $s_{\tilde{i}\tilde{h}}$, $i = y, 1, \ldots, p$,
- σ_{hh} with its consistent estimator $s_{\tilde{h}\tilde{h}}$,
- $\sigma_{\delta_i\delta_j}$, $i, j = y, 1, \ldots, p$, with their consistent estimators (see Equation (A.21) in Schmid 2007)

$$\sigma_{\tilde{\delta}_i\tilde{\delta}_{j,c}} := K\left(s_{\tilde{i}\tilde{j}} - \frac{s_{\tilde{i}\tilde{h}}s_{\tilde{j}\tilde{h}}}{s_{\tilde{h}\tilde{h}}}\right) \quad i,j = y, 1, \ldots, p \tag{31}$$

- σ_{ij}, $i, j = \mathrm{y}, 1, \ldots, p$, with their consistent estimators (see (11) and (12))

$$\sigma_{\tilde{i}\tilde{j},c} := Ks_{\tilde{i}\tilde{j}} - (K-1)\frac{s_{\tilde{i}\tilde{h}}s_{\tilde{j}\tilde{h}}}{s_{\tilde{h}\tilde{h}}} \quad i,j = y, 1, \ldots, p \tag{32}$$

- $\Sigma_{\tilde{x}\tilde{x}}$ *with* $S_{\tilde{x}\tilde{x}}$
- $\sigma_{\tilde{x}\tilde{y}}$ *with* $s_{\tilde{x}\tilde{y}}$

5. Finite Sample Behavior of $\tilde{b}_c$

In this section we check whether the asymptotic results derived in Sections 3 and 4 hold in realistic data situations. To this end, a simulation study was carried out using the statistical software R, version 2.7.0 (R Development Core Team 2008). The model we studied was a linear regression with two normally distributed regressors X_1 and X_2. The variance parameters were $\sigma_{11} = 1$, $\sigma_{22} = 4$, and $\sigma_{12} = 1$, which corresponds to a correlation of 0.5 between the two regressors.

5.1. Bias of $\tilde{b}_c$ for Finite Samples

To study the bias of $\tilde{b}_c$, we took $K = 3$ (which is the group size commonly used in practice) and $\beta_0 = 0$. For simplicity, we kept $\beta_2 = -1$ fixed. The residual variance σ_ε^2 was set to 9, which is a rather large value if compared to the values of $\sigma_{11} = 1$ and $\sigma_{22} = 4$.

Now, for various values of β_1, we estimated the bias of $\tilde{b} = (\tilde{b}_1, \tilde{b}_2)'$ and $\tilde{b}_c = (\tilde{b}_{1,c}, \tilde{b}_{2,c})'$ from 1,000 randomly generated data sets $(x_{\nu 1}, x_{\nu 2}, y_\nu)$, $\nu = 1, \ldots, n$. The sorting variables we used were (1) the first principal component projection using the empirical correlation matrix of (Y, X_1, X_2), (2) the sum of z-scores using the empirical variances of Y, X_1, and X_2, (3) the dependent variable Y, and (4) the regressor X_1.

In Figures 1 to 4, bias($\tilde{b}_1$) and bias($\tilde{b}_{1,c}$) are plotted vs. β_1 for $n = 150$ and $n = 600$. Obviously, the finite sample bias of $\tilde{b}_{1,c}$ is close to zero if $n \geq 150$. Moreover, it can be seen from Figures 1 and 3 that the bias of $\tilde{b}_1$ does not converge to 0 as n increases. As expected, the only exception is the case where X_1 is the sorting variable (since in this case

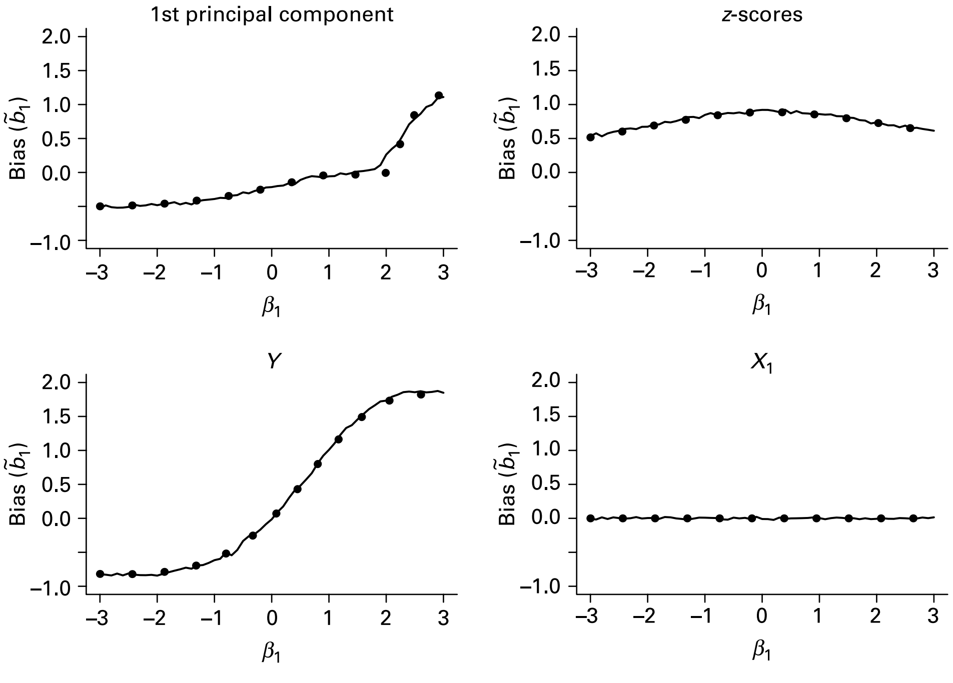

Fig. 1. Bias of $\tilde{b}_1$ for various sorting variables and $n = 150$. The dotted lines correspond to the true asymptotic bias curves

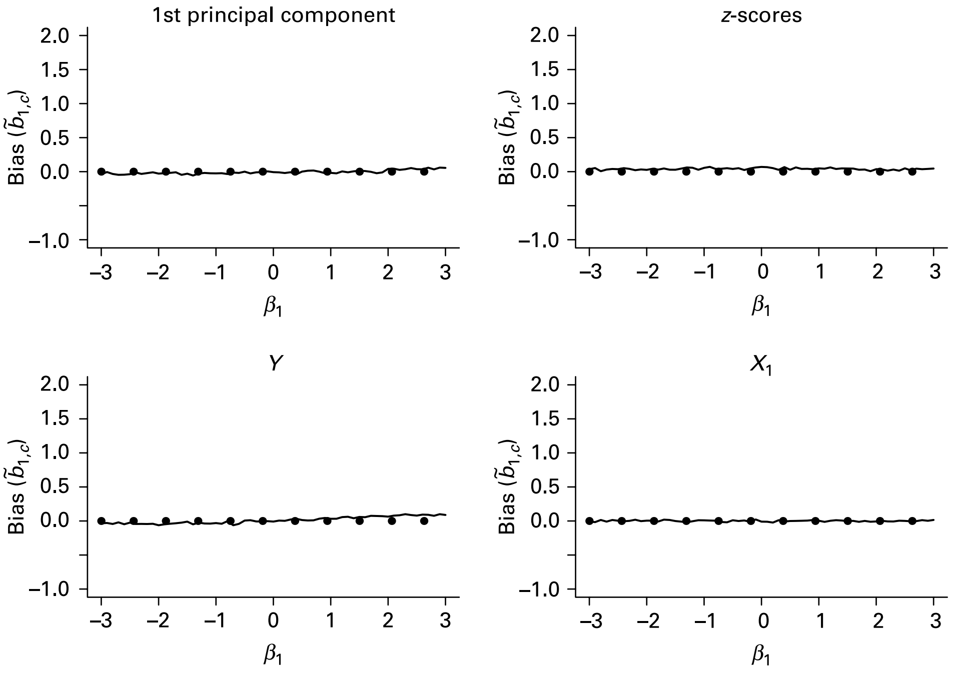

Fig. 2. Bias of $\tilde{b}_{1,c}$ for various sorting variables and $n = 150$. The dotted lines correspond to the true asymptotic bias curves (which are equal to zero since $\tilde{b}_{1,c}$ is a consistent estimator of β_1)

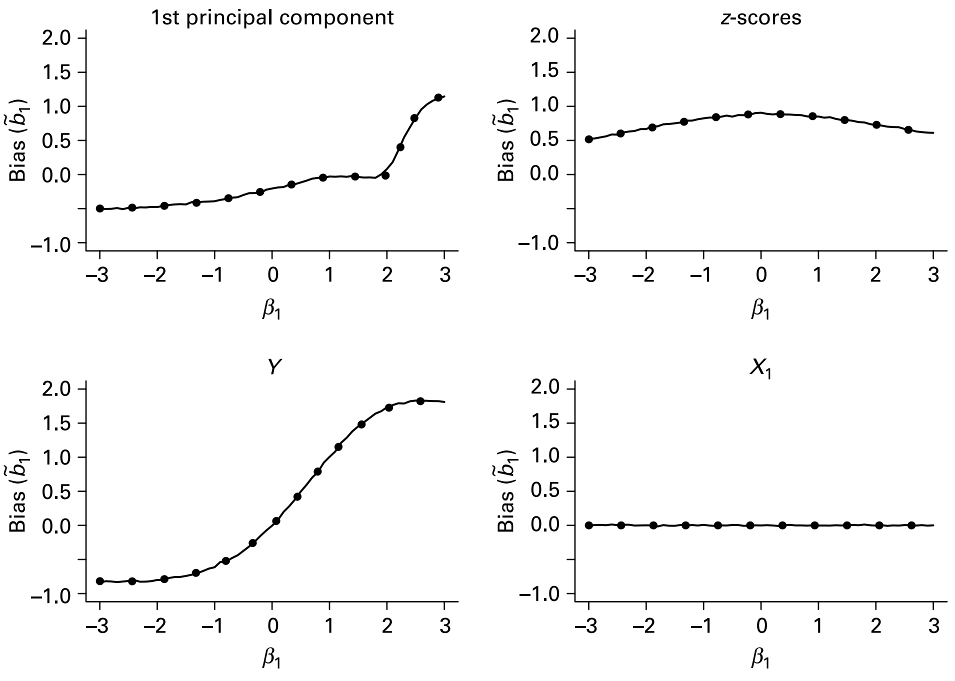

Fig. 3. Bias of $\tilde{b}_1$ for various sorting variables and $n = 600$. The dotted lines correspond to the true asymptotic bias curves

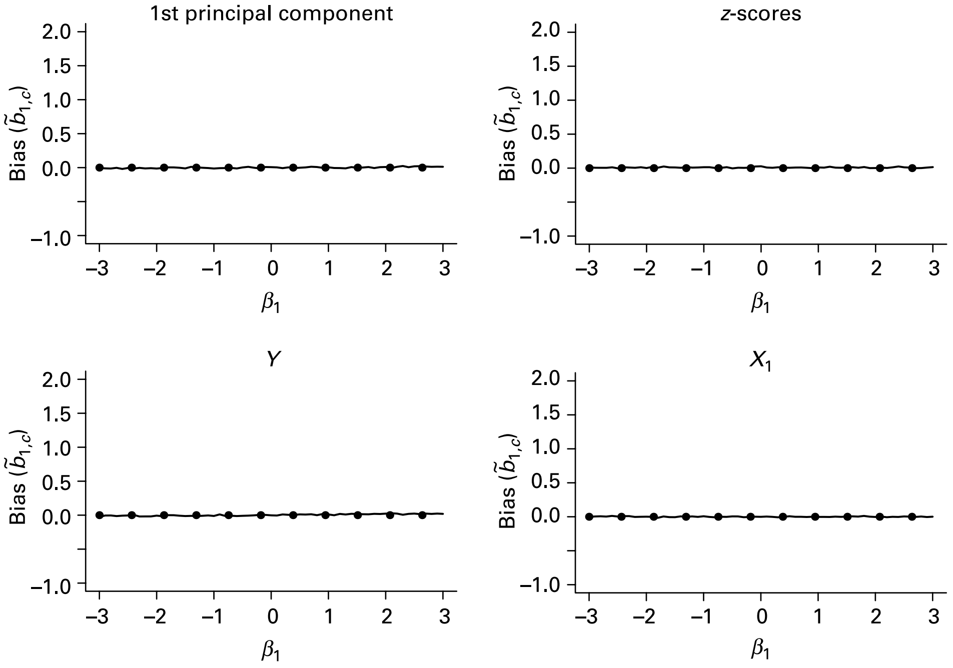

Fig. 4. Bias of $\tilde{b}_{1,c}$ for various sorting variables and n = 600. The dotted lines correspond to the true asymptotic bias curves (which are equal to zero since $\tilde{b}_{1,c}$ is a consistent estimator of β_1)

$\tilde{b}_1$ is a consistent estimator of β_1). The estimators $\tilde{b}_2$ and $\tilde{b}_{2,c}$ show a similar behavior as $\tilde{b}_1$ and $\tilde{b}_{1,c}$, respectively, but are omitted here due to space limitations.

5.2. *Variance of $\tilde{b}_c$ for Finite Samples*

Figures 5 and 6 contain the variances of $\sqrt{n}\tilde{b}_{1,c}$, which were estimated from the simulated data for $n = 150$ and $n = 600$. Moreover, Figures 5 and 6 show the averages of the estimated asymptotic variances of $\sqrt{n}\tilde{b}_{1,c}$, as well as the corresponding true asymptotic variances. We see that if the sample size is small ($n = 150$), var($\tilde{b}_{1,c}$) is underestimated by its asymptotic counterpart. For larger sample sizes ($n = 600$) the asymptotic variance of $\tilde{b}_{1,c}$ is a good approximation of the true variance of $\tilde{b}_{1,c}$. The variance of $\tilde{b}_{2,c}$ and the covariance of $\tilde{b}_{1,c}$ and $\tilde{b}_{2,c}$ show a similar behavior as the variance of $\tilde{b}_{1,c}$ but are omitted here due to space limitations.

6. Summary and Conclusion

We have analyzed the effect of single-axis sorting microaggregation on the least squares estimation of a linear regression model in continuous variables. In Section 3 we have shown that the LS estimators of the linear model (computed from the microaggregated data) are not necessarily consistent estimators of the true model parameters. It is only in the special case where the sorting variable is a linear combination of the regressors that the LS estimator of the slope parameter vector β turns out to be consistent. Although aggregating with respect to a linear combination of the regressors therefore seems to be

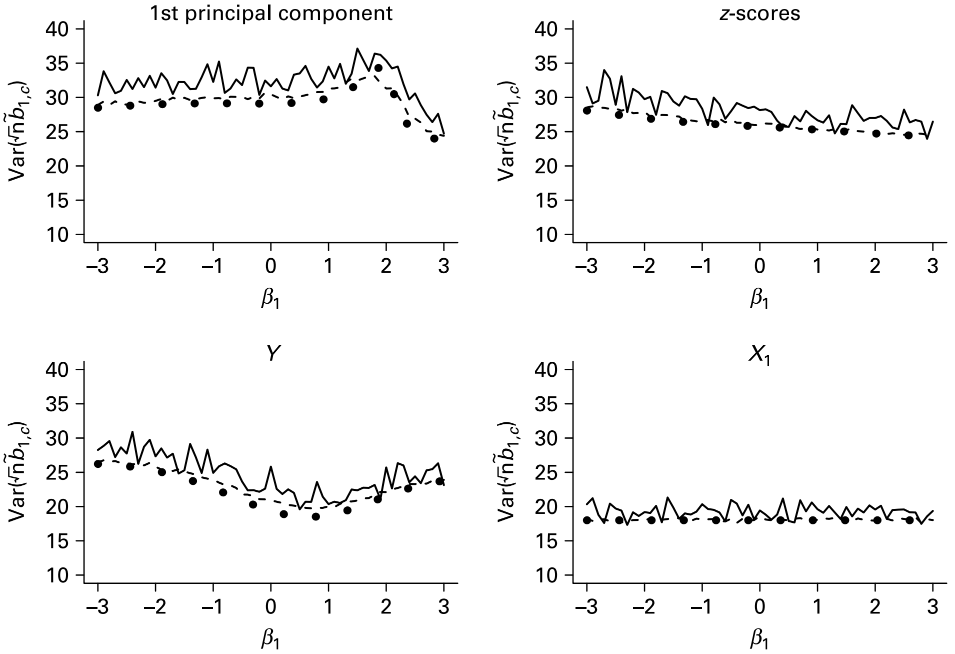

Fig. 5. Variance of $\sqrt{n}\tilde{b}_{1,c}$, as estimated from simulation (solid lines: true variances, dashed lines: averages of the estimated asymptotic variances, dotted lines: true asymptotic variances, $n = 150$)

more convenient for statistical analysis, it has to be pointed out that data holders usually do not know *before* anonymization which variables will later serve as the regressors in a linear model. Thus, investigating microaggregation with respect to an arbitrary sorting variable H is a relevant case.

The main result of the article is the development of a corrected estimator that removes the aggregation bias of the LS estimator of β. We also derived the asymptotic covariance matrix of the corrected estimator. The simulation study in Section 5 has shown that the correction procedure already works well if the sample size is moderately high ($n \geq 150$). It should be noted, however, that the finite sample behavior of the corrected estimator $\tilde{b}_c$ also depends on the number of covariates in the linear model ("curse of dimensionality," see Aggarwal 2005). For the simulation study $p = 2$ covariates were used. Further empirical work conducted by Schmid (2007) suggests that the (finite sample) bias and variance of $\tilde{b}_c$ are likely to increase as p gets larger. However, for any value of p, $\tilde{b}_c$ seems to perform better than the LS estimator $\tilde{b}$. From a numerical point of view it is possible that the computation of the corrected estimator may cause problems in some situations, since it cannot be ruled out completely that the denominator in (10) becomes close to zero. When computing the corrected estimator from our real-world and simulated data sets, however, no numerical problems have ever been encountered.

It should further be noted that, in order to prove the results presented in this article, we assumed the variables in the linear model and the sorting variable to be jointly normally distributed. This assumption was needed to guarantee the existence of the inverse linear regression Models (4) and (5), where the independence of the residuals $\delta_y, \delta_1, \ldots, \delta_p$

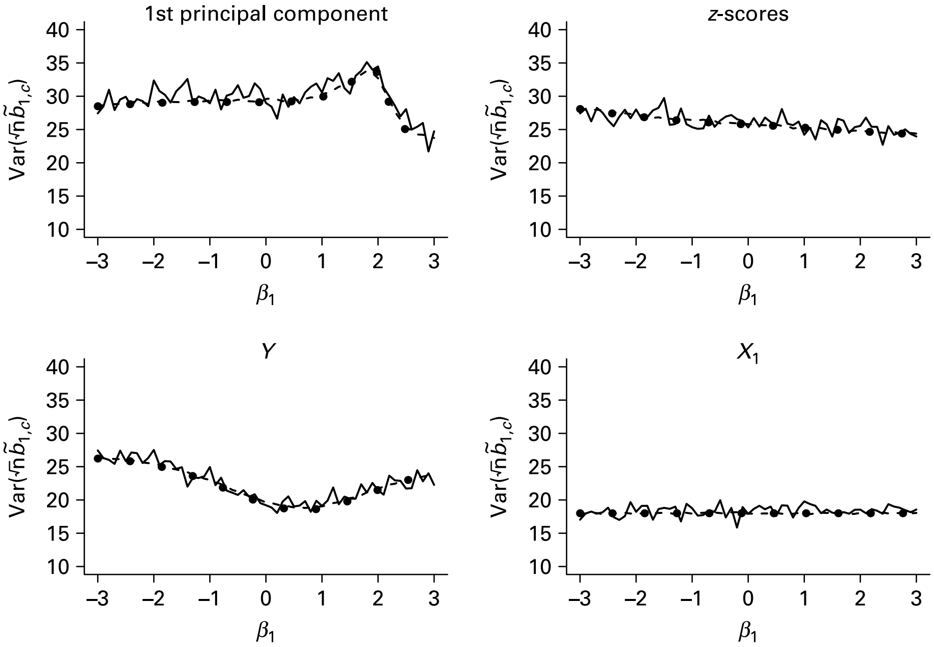

Fig. 6. Variance of $\sqrt{n}\tilde{b}_{1,c}$, as estimated from simulation (solid lines: true variances, dashed lines: averages of the estimated asymptotic variances, dotted lines: true asymptotic variances, $n = 600$)

and the sorting variable H was used for deriving the probability limits of the empirical first and second moments computed from the microaggregated data. Clearly, in many data situations, the normality assumption is unlikely to hold. For this reason, an analysis of the robustness of $\tilde{b}_c$ against deviations from the normality assumption is needed. Empirical studies presented by Schmid et al. (2007) and Schmid (2007) suggest that skewed and fat-tailed distributions can indeed lead to an asymptotic bias of the corrected estimator $\tilde{b}_c$. However, small or moderate deviations from normality do not seem to have a large effect on the behavior of $\tilde{b}_c$.

In order to enable data users to carry out the estimation procedure developed in Sections 3 and 4, data holders are required to provide the aggregated data values of the sorting variable. In most cases, this requirement will not severely affect the disclosure risk of a data set, as the K-anonymity of the data will still be guaranteed. If the sorting variable is a linear combination of the variables in the linear model, data holders only have to provide the coefficients $c_y, c_1, \ldots, c_p$ of this combination. Data users are then able to reconstruct the aggregated data values of H from $c_y, c_1, \ldots, c_p$ and from the microaggregated data. Moreover, there are special types of sorting variables where the coefficients $c_y, c_1, \ldots, c_p$ can be consistently estimated from the microaggregated data (see Schmid 2007, Section 4.3.3). Simulations presented in Schmid (2007) suggest that the additional variance induced by the estimation of $c_y, c_1, \ldots, c_p$ is negligible.

A limitation of the correction procedure presented in this article is that it only applies to a linear model with *continuous* regressors. If discrete regressors had been included in the linear model, the following problems would have occurred: (1) As mentioned in Section 2, discrete variables are usually not microaggregated but anonymized by means of other disclosure control techniques. As a consequence, we would not have been able to analyze the effect of *microaggregation* on the LS estimators, but a mixture of effects caused by *both* microaggregation *and* the disclosure control techniques for the discrete data. (2) With discrete regressors in the model equation, additional assumptions regarding the distribution of the discrete variables would have been necessary. With these assumptions, the data would not have been normally distributed any more. As the normality assumption plays a key role in the proofs of the lemmas in Sections 3 and 4, it will be challenging to develop a correction procedure for $\tilde{b}$ in case of a linear model with both discrete and continuous regressors.

It should finally be pointed out that single-axis sorting techniques are related to a clearly specified mathematical model which is based on the theory of order statistics. This model made it possible to carry out the analytical investigation in Sections 3 and 4. The methods presented in this article can therefore be regarded as a contribution to the analytical understanding of microaggregation by single-axis sorting and its effect on statistical estimation techniques. However, it is not guaranteed that the correction techniques developed in the article will also work if other microaggregation techniques (such as iterative techniques with variable-sized groups) are applied to a data set, since the mathematical assumptions made in this article may not hold if other microaggregation techniques than single-axis sorting are used. Specifying appropriate mathematical assumptions for microaggregation techniques with variable-sized groups will therefore be the basis of an investigation of the effect of these techniques on statistical model estimation.

7. References

Aggarwal, C.C. (2005). On k-anonymity and the Curse of Dimensionality. In Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, Norway, 901–909.

Defays, D. and Anwar, M.N. (1998). Masking Microdata Using Microaggregation. Journal of Official Statistics, 14, 449–461.

Defays, D. and Nanopoulos, P. (1993). Panels of Enterprises and Confidentiality: The Small Aggregates Method. In Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys. Ottawa: Statistics Canada, 195–204.

Dhrymes, P.J. (1984). Mathematics for Econometrics, (Second Edition). New York: Springer.

Domingo-Ferrer, J., Martinez-Balleste, A., Mateo-Sanz, J.M., and Sebe, F. (2006). Efficient Multivariate Data-oriented Microaggregation. The VLDB Journal, 15, 355–369.

Domingo-Ferrer, J. and Mateo-Sanz, J.M. (2002). Practical Data-oriented Microaggregation for Statistical Disclosure Control. IEEE Transactions on Knowledge and Data Engineering, 14, 189–201.

Domingo-Ferrer, J., Sebe, F., and Solanas, A. (2008). A Polynomial-time Approximation to Optimal Multivariate Microaggregation. Computers and Mathematics with Applications, 55, 714–732.

Domingo-Ferrer, J. and Torra, V. (2005). Ordinal, Continuous and Heterogeneous k-anonymity Through Microaggregation. Data Mining and Knowledge Discovery, 11, 195–212.

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (2001). Confidentiality, Disclosure, and Data Access. Amsterdam: North-Holland.

Evans, M., Hastings, N., and Peacock, B. (1993). Statistical Distributions, (Second edition). New York: Wiley.

Feige, E.L. and Watts, H.W. (1972). An Investigation of the Consequences of Partial Aggregation of Micro-economic Data. Econometrica, 40, 343–360.

Hansen, S.L. and Mukherjee, S. (2003). A Polynomial Algorithm for Optimal Univariate Microaggregation. IEEE Transactions on Knowledge and Data Engineering, 15, 1043–1044.

Laszlo, M. and Mukherjee, S. (2005). Minimum Spanning Tree Partitioning Algorithm for Microaggregation. IEEE Transactions on Knowledge and Data Engineering, 17, 902–911.

Lenz, R. (2006). Measuring the Disclosure Protection of Micro Aggregated Business Microdata. An Analysis Taking as an Example the German Structure of Costs Survey. Journal of Official Statistics, 22, 681–710.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE '07), Istanbul, Turkey.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramanian, M. (2006). l-Diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE '06), Atlanta, GA, U.S.A.

Mateo-Sanz, J.M. and Domingo-Ferrer, J. (1998). A Comparative Study of Microaggregation Methods. Questiio, 22, 511–526.

Oganian, A. and Domingo-Ferrer, J. (2001). On the Complexity of Optimal Microaggregation for Statistical Disclosure Control. Statistical Journal of the United Nations Economic Commission for Europe, 18, 345–354.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M., and Vorgrimler, D. (2005). Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, Volume 4 of Statistik und Wissenschaft. Wiesbaden: Statistisches Bundesamt [In German].

Samarati, P. (2001). Protecting Respondents' Identities in Microdata Release. IEEE Transactions on Knowledge and Data Engineering, 13, 1010–1027.

Schmid, M. (2007). Estimation of a Linear Regression With Microaggregated Data. Munich: Verlag Dr. Hut, Dissertation, University of Munich.

Schmid, M. and Schneeweiss, H. (2005). The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study. In W. Pohlmeier, G. Ronning, and J. Wagner (eds), Econometrics of Anonymized Micro Data, Volume 225 of Jahrbücher für Nationalökonomie und Statistik, 529–543. Stuttgart: Lucius & Lucius.

Schmid, M., Schneeweiss, H., and Küchenhoff, H. (2007). Estimation of a Linear Regression Under Microaggregation With the Response Variable as a Sorting Variable. Statistica Neerlandica, 61, 407–431.

Solanas, A. and Martinez-Balleste, A. (2006). V-MDAV: A Multivariate Microaggregation with Variable Group Size. In Proceedings in Computational Statistics (COMPSTAT 2006), Rome, Italy, 917–925. Berlin: Physica-Verlag.

Solanas, A., Martinez-Balleste, A., Mateo-Sanz, J.M., and Domingo-Ferrer, J. (2006). Multivariate Microaggregation Based on Genetic Algorithms. In Proceedings of the 3rd IEEE Conference on Intelligent Systems (IEEE IS' 2006). New York: IEEE Press.

Sweeney, L. (2002). k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10, 557–570.

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

Received October 2007
Revised August 2008