Schmid, Matthias

Working Paper

The effect of single-axis sorting on the estimation of a linear regression

Discussion Paper, No. 472

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 (2006) : The effect of single-axis sorting on the estimation of a linear regression, Discussion Paper, No. 472, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München, https://doi.org/10.5282/ubm/epub.1840

This Version is available at:
https://hdl.handle.net/10419/31097

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

Matthias Schmid

Department of Statistics, University of Munich
Ludwigstr. 33, 80539 Munich, Germany
matthias.schmid@stat.uni-muenchen.de

Abstract

Microaggregation is one of the most important statistical disclosure control techniques for continuous data. The basic principle of microaggregation is to group the observations in a data set and to replace them by their corresponding group means. In this paper, we consider single-axis sorting, a frequently applied microaggregation technique where the formation of groups depends on the magnitude of a sorting variable related to the variables in the data set. The paper deals with the impact of this technique on a linear model in continuous variables. We show that parameter estimates are asymptotically biased if the sorting variable depends on the response variable of the linear model. 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.

Keywords: Asymptotic variance, consistent estimation, disclosure control, linear model, microaggregation, sorting variable.

1 Introduction

Microaggregation is one of the most frequently applied statistical disclosure control techniques for continuous microdata (Defays and Nanopoulos (1993), Domingo-Ferrer and Mateo-Sanz (2002)). The main idea of microaggregation is to subdivide the observations in a data set into small groups (with minimum group size A) and to replace the original data values by their corresponding group means. Thus, as each observation in the microaggregated data set appears at least A times, individual records cannot be identified, and the disclosure risk of the anonymized data is kept low. However, microaggregation can severely affect the results of statistical analyses (see, e.g., Statistisches Bundesamt (2005)).

In a previous paper, Schmid and Schneeweiss (2005) have analyzed the effect

of microaggregation on the least squares (LS) estimation of a linear regression in continuous variables. To aggregate the data, Schmid and Schneeweiss have used the dependent variable in the linear model as a sorting variable, thus obtaining groups consisting of observations that have similar values for the sorting variable. As Schmid and Schneeweiss have shown, the naive LS estimators of a linear model are asymptotically biased in this case. By taking the bias into account, Schmid and Schneeweiss have also developed a consistent estimator for the model parameters.

The present paper generalizes the results of Schmid and Schneeweiss (2005) to the case where an arbitrary sorting variable H is used for microaggregation. In the literature, this technique has been referred to as *single-axis sorting* microaggregation. We assume that the variables of the linear model and the sorting variable are jointly normally distributed. Consequently, H does not have to be the dependent variable or one of the regressors, but can also be an arbitrary linear combination of the variables in the linear model (such as the first principal component projection or the sum of z-scores).

In the following, we will derive *analytically* the asymptotic properties of the naive LS estimators when applied to data that have been microaggregated with respect to H. We will not only determine the (asymptotic) bias, but also develop a new estimation procedure that corrects for the bias, leading to a consistent estimator of the linear model. In addition, the asymptotic covariance matrix of the corrected LS estimator of the slope parameter vector β will be derived.

Section 2 starts with a brief description of single-axis sorting microaggregation. In Section 3 we derive theoretical results on the effects of this procedure on the 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 of the slope parameter vector. Section 5 contains a simulation study on the results derived in Sections 3 and 4. In Section 6 we apply our results to the 2003 Munich Rent Data. Section 7 deals with the estimation of the aggregated values of H from the anonymized data. The results of this paper are summarized in Section 8.

2 Microaggregation by a Sorting Variable

We consider microaggregation with respect to a sorting variable in the data set. The microaggregation procedure is as follows: First, the data set has to be ordered according to the magnitude of the sorting variable. After having chosen a fixed group size A, the sorted data set is subdivided into small groups, each consisting of A adjacent data values. For simplicity, we assume that the sample size n is a multiple of A. In each of the n/A groups the data are averaged, and the averages are assigned to the items of the group.

For example, consider a data set with 6 observations and 3 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

.

The first principal component values of x_1, x_2, and y are given by $h = (-0.17, 0.09, 0.24, 0.97, -0.59, -0.54)$. Now, if the first principal component projection is used as a sorting variable and $A = 3$, we obtain 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})	(-0.59)	(-0.54)	(-0.17)	(0.09)	(0.24)	(0.97)

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

3.1 Notation

Consider the multiple linear regression model

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

Y denotes the response (or endogenous) variable, while $X_1, \ldots, X_p$ denote the covariates (or exogenous variables). Further consider a random variable H, which will later serve as the sorting variable. Y, $X_1, \ldots, X_p$, and H are assumed to be jointly normally distributed random variables with variances σ_{yy}, $\sigma_{11}, \ldots, \sigma_{pp}$, σ_{hh}. The random error ϵ is assumed to be independent of $(X_1, \ldots, X_p)$ with zero mean and variance σ_ϵ^2. The objective is to estimate the parameter vector $(\beta_0, \beta_1, \ldots, \beta_p)'$ and the residual variance σ_ϵ^2 from an i.i.d. sample $(y_z, x_{z1}, \ldots, x_{zp})$, $z = 1, \ldots, n$. Let $y := (y_1, \ldots, y_n)'$ and $x_i := (x_{1i}, \ldots, x_{ni})'$, $i = 1, \ldots, p$, contain the data values. Let $h := (h_1, \ldots, h_n)'$ contain the data values of H. The vectors containing the aggregated data are denoted by $\tilde{y}$, $\tilde{x}_1, \ldots, \tilde{x}_p$, and $\tilde{h}$. For simplicity, it is assumed throughout that n is a multiple of A. Note that in this case, 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. We denote the covariance of X_i and X_j by σ_{ij}, $i, j = 1, \ldots, p$, the covariance of X_i and Y by σ_{iy}, $i = 1, \ldots, p$, the covariance of X_i and H by σ_{ih}, $i = 1, \ldots, p$, and the covariance of Y and H by σ_{yh}.

Further denote the empirical covariance of x_i and x_j by s_{ij} and the empirical covariance of $\tilde{x}_i$ and $\tilde{x}_j$ by $\tilde{s}_{ij}$:

$$s_{ij} := \frac{1}{n} \sum_{z=1}^{n} (x_{zi} - \bar{x}_i)(x_{zj} - \bar{x}_j) \ , \ i, j = 1, \ldots, p \, , \tag{2}$$

$$\tilde{s}_{ij} := \frac{1}{n} \sum_{z=1}^{n} (\tilde{x}_{zi} - \bar{x}_i)(\tilde{x}_{zj} - \bar{x}_j) \ , \ i, j = 1, \ldots, p \, . \tag{3}$$

The covariance matrix of $(X_1, \ldots, X_p)$ is denoted by $\Sigma := (\sigma_{ij})_{i,j=1,\ldots,p}$. Similarly let $\sigma_{xy} := (\sigma_{iy})_{i=1,\ldots,p}$ and $\sigma_{xh} := (\sigma_{ih})_{i=1,\ldots,p}$ be the covariance (column)

vectors of $(X_1, \ldots, X_p)$ and Y and of $(X_1, \ldots, X_p)$ and H, respectively. The empirical variances of y, $\tilde{y}$, h, and $\tilde{h}$ are denoted by s_{yy}, $\tilde{s}_{yy}$, s_{hh}, and $\tilde{s}_{hh}$, respectively, and the empirical covariances of x_i and y, $\tilde{x}_i$ and $\tilde{y}$, x_i and h, and $\tilde{x}_i$ and $\tilde{h}$ are denoted by s_{iy}, $\tilde{s}_{iy}$, s_{ih}, and $\tilde{s}_{ih}$, respectively. Finally let

$$s_{xy} := \begin{pmatrix} s_{1y} \\ \vdots \\ s_{py} \end{pmatrix} \quad , \quad \tilde{s}_{xy} := \begin{pmatrix} \tilde{s}_{1y} \\ \vdots \\ \tilde{s}_{py} \end{pmatrix} \quad , \quad i = 1, \ldots, p, \tag{4}$$

$$s_{xh} := \begin{pmatrix} s_{1h} \\ \vdots \\ s_{ph} \end{pmatrix} \quad , \quad \tilde{s}_{xh} := \begin{pmatrix} \tilde{s}_{1h} \\ \vdots \\ \tilde{s}_{ph} \end{pmatrix} \quad , \quad i = 1, \ldots, p, \tag{5}$$

and let $S := (s_{ij})_{i,j=1,\ldots,p}$ and $\tilde{S} := (\tilde{s}_{ij})_{i,j=1,\ldots,p}$ be the empirical covariance matrices of $(x_1, \ldots, x_p)$ and $(\tilde{x}_1, \ldots, \tilde{x}_p)$, respectively.

3.2 Examples of Sorting Variables

It follows from the joint normality of $Y, X_1, \ldots, X_p$, and H that

$$H = c_y Y + c_1 X_1 + \cdots + c_p X_p + \varphi \ , \tag{6}$$

where $c := (c_y, c_1, \ldots, c_p)'$ is a vector of coefficients and φ is a normally distributed random variable with zero mean and variance σ_φ^2.

Usually, the sorting variable is an exact linear combination of the variables in the linear model, implying that $\varphi \equiv 0$. Popular choices for the sorting variable include

- the dependent variable Y: In this case, $\varphi \equiv 0$, $c_y = 1$, $c_1 = \cdots = c_p = 0$.
- a regressor X_i: In this case, $\varphi \equiv 0$, $c_y = 0$, $c_1 = \cdots = c_{i-1} = 0$, $c_i = 1$, $c_{i+1} = \cdots = c_p = 0$.
- the first principal component projection of $Y, X_1, \ldots, X_p$: In this case, $\varphi \equiv 0$, and c is the eigenvector associated to the largest eigenvalue of the covariance or correlation matrix of $Y, X_1, \ldots, X_p$.

- the sum of z-scores of the variables in the linear model: In this case $\varphi \equiv 0$, $c_y = \sigma_{yy}^{-1/2}$, $c_1 = \sigma_{11}^{-1/2}$,...,$c_p = \sigma_{pp}^{-1/2}$.

In the latter two cases, the coefficients $c_y, c_1, \ldots, c_p$ are not known to the data providers and thus have to be estimated from the non-aggregated data.

Note that H can also be a variable of the originally collected data set, which however is not used in the regression analysis. In this case, typically $\varphi \not\equiv 0$.

3.3 Consistent Estimation of β

We focus on the estimation of the vector of genuine regression coefficients $\beta := (\beta_1, \ldots, \beta_p)'$. 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 naive least squares estimator of β by $\tilde{b}$, which is given by

$$\tilde{b} := \tilde{S}^{-1}\tilde{s}_{xy} \ . \tag{7}$$

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

Lemma 1. *Consider the inverse linear relationships*

$$X_i = \alpha_i + \gamma_i H + \delta_i \ , \ i = 1, \ldots, p \ , \tag{8}$$
$$Y = \alpha_y + \gamma_y H + \delta_y \ , \tag{9}$$

which exist due to the joint normality of $Y, X_1, \ldots, X_p$, and H. The δ_i's, $i = y, 1, \ldots, p$, are random variables, independent of H, with zero mean and variances and covariances $\sigma_{\delta_i\delta_j}$, $i, j = y, 1, \ldots, p$. The following probability limits exist:

a) $\text{plim}_{n\to\infty}\tilde{s}_{hh} = \sigma_{hh}$,

b) $\text{plim}_{n\to\infty}\tilde{s}_{xh} = \sigma_{xh}$,

c) $\text{plim}_{n\to\infty}\tilde{s}_{yh} = \sigma_{yh}$,

d) $\text{plim}_{n\to\infty}\tilde{S} = \frac{1}{A}\Sigma + \left(1 - \frac{1}{A}\right)\frac{\sigma_{xh}\sigma_{xh}'}{\sigma_{hh}} =: \tilde{\Sigma}$,

e) $\text{plim}_{n\to\infty}\tilde{s}_{xy} = \frac{1}{A}\sigma_{xy} + \left(1 - \frac{1}{A}\right)\frac{\sigma_{yh}}{\sigma_{hh}}\sigma_{xh} =: \tilde{\sigma}_{xy}$.

Proof: As H is a sorting variable related to $Y, X_1, \ldots, X_p$ by the linear regression models (8) and (9), Lemma 1 directly follows from Lemma 1 in Schmid and Schneeweiss (2005).

With Lemma 1 and setting $a := (A-1)/\sigma_{hh}$, the probability limit of $\tilde{b}$ can be evaluated as

$$\begin{aligned}
\tilde{\beta} &:= \operatorname{plim}_{n\to\infty} \tilde{b} = \tilde{\Sigma}^{-1}\tilde{\sigma}_{xy} \\
&= \Big(\Sigma + a\sigma_{xh}\sigma_{xh}'\Big)^{-1}\Big(\sigma_{xy} + a\sigma_{xh}\sigma_{yh}\Big) \\
&= \Big(\Sigma^{-1} - \Sigma^{-1}\frac{a\sigma_{xh}\sigma_{xh}'}{1 + a\sigma_{xh}'\Sigma^{-1}\sigma_{xh}}\Sigma^{-1}\Big)\Big(\sigma_{xy} + a\sigma_{xh}\sigma_{yh}\Big) \ . \qquad (10)
\end{aligned}$$

In order to obtain (10), we used a matrix inversion formula which can be found, e.g., in Dhrymes (1984), Corollary 5. Using $\beta = \Sigma^{-1}\sigma_{xy}$, it follows from (10) that

$$\begin{aligned}
\tilde{\beta} &= \beta + a\sigma_{yh}\Sigma^{-1}\sigma_{xh} - \frac{a\sigma_{xh}'\Sigma^{-1}\sigma_{xy}}{1 + a\sigma_{xh}'\Sigma^{-1}\sigma_{xh}}\,\Sigma^{-1}\sigma_{xh} \\
&\quad - \frac{a^2\sigma_{yh}\sigma_{xh}'\Sigma^{-1}\sigma_{xh}}{1 + a\sigma_{xh}'\Sigma^{-1}\sigma_{xh}}\,\Sigma^{-1}\sigma_{xh} \\
&= \beta + \frac{a(\sigma_{yh} - \sigma_{xh}'\Sigma^{-1}\sigma_{xy})}{1 + a\sigma_{xh}'\Sigma^{-1}\sigma_{xh}}\,\Sigma^{-1}\sigma_{xh} \ . \qquad (11)
\end{aligned}$$

It is easily seen from (11) that the naive LS estimator $\tilde{b}$ is asymptotically biased. In case of the non-aggregated data (i.e. $A = 1$), the asymptotic bias is equal to 0.

In the special case where H is an exact linear combination of $Y, X_1, \ldots, X_p$ (see Section 3.2), we obtain $\sigma_{yh} - \sigma_{xh}'\Sigma^{-1}\sigma_{xy} = c_y(\sigma_{yy} - \sigma_{xy}'\Sigma^{-1}\sigma_{xy})$. Thus (11) becomes

$$\tilde{\beta} = \beta + \frac{ac_y(\sigma_{yy} - \sigma_{xy}'\Sigma^{-1}\sigma_{xy})}{1 + a\sigma_{xh}'\Sigma^{-1}\sigma_{xh}}\,\Sigma^{-1}\sigma_{xh} \ . \qquad (12)$$

From (12) we see that $\tilde{b}$ is a consistent estimator of β as long as $c_y = 0$. This is the case when a particular regressor or a linear combination of the regressors serves as the sorting variable.

If Y is the sorting variable ($c_y = 1$), we have $\sigma_{xh} = \sigma_{xy}$ and $\sigma_{hh} = \sigma_{yy}$. With a little algebra, it can be shown that in this case

$$\tilde{\beta} \;=\; \frac{A}{1 + (A-1)\sigma'_{xy}\Sigma^{-1}\sigma_{xy}/\sigma_{yy}}\;\beta\;. \tag{13}$$

This is the same relationship as the one derived in Schmid and Schneeweiss (2005). We see that if Y is the sorting variable, $\tilde{b}$ is asymptotically biased, with $\tilde{\beta}$ being proportional to β.

In order to construct a consistent estimator of β, we start from $\beta = \Sigma^{-1}\sigma_{xy}$ and replace Σ with

$$\Sigma \;=\; \left(A\tilde{\Sigma} - (A-1)\,\frac{\sigma_{xh}\sigma'_{xh}}{\sigma_{hh}}\right) \tag{14}$$

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

$$\sigma_{xy} \;=\; \left(A\tilde{\sigma}_{xy} - (A-1)\frac{\sigma_{yh}}{\sigma_{hh}}\sigma_{xh}\right) \tag{15}$$

from Lemma 1e). By algebraic manipulations similar to those that led to (11), this yields

$$\beta \;=\; \tilde{\beta} + \frac{(A-1)(\sigma'_{xh}\tilde{\Sigma}^{-1}\tilde{\sigma}_{xy} - \sigma_{yh})}{A\sigma_{hh} - (A-1)\sigma'_{xh}\tilde{\Sigma}^{-1}\sigma_{xh}}\;\tilde{\Sigma}^{-1}\sigma_{xh}\;, \tag{16}$$

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

$$\tilde{b}_c \;:=\; \tilde{b} + \frac{(A-1)(\tilde{s}'_{xh}\tilde{S}^{-1}\tilde{s}_{xy} - \tilde{s}_{yh})}{A\tilde{s}_{hh} - (A-1)\tilde{s}'_{xh}\tilde{S}^{-1}\tilde{s}_{xh}}\;\tilde{S}^{-1}\tilde{s}_{xh}\;. \tag{17}$$

Note that the computation of (17) requires the aggregated data values of the sorting variable H to be known to the data user. This either implies that the data holder provides the aggregated data values of H or that $\tilde{h}$ can be reconstructed from the aggregated data values $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$ (see Section 7). With a little algebra, it can also be shown that if $c_y = 0$ and $\varphi \equiv 0$, the expression $\tilde{s}'_{xh}\tilde{S}^{-1}\tilde{s}_{xy} - \tilde{s}_{yh}$ in (17) is equal to 0. Thus, in this case, the (consistent) naive estimator $\tilde{b}$ is equal to the corrected estimator $\tilde{b}_c$.

A consistent estimator of the intercept β_0 is simply given by

$$\tilde{b}_{0c} \;:=\; \bar{\tilde{y}} - (\tilde{b}_{1c}\bar{\tilde{x}}_1 + \cdots + \tilde{b}_{pc}\bar{\tilde{x}}_p) \;, \tag{18}$$

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

Furthermore, from (14) and (17), we obtain a consistent estimator of the residual variance $\sigma_\epsilon^2 = \sigma_{yy} - \beta'\Sigma\beta$:

$$\sigma^2_{\epsilon,c} := \left(A\tilde{s}_{yy} - (A-1)\frac{\tilde{s}^2_{yh}}{\tilde{s}_{hh}}\right) - \tilde{\beta}'_c\left(A\tilde{S} - (A-1)\frac{\tilde{s}_{xh}\tilde{s}'_{xh}}{s_{hh}}\right)\tilde{\beta}_c \;. \tag{19}$$

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

To derive the asymptotic covariance matrix of $\tilde{b}_c$, we will use the following notation:

- Two random sequences a_n and b_n are said to be "asymptotically equivalent", written $a_n \sim b_n$, if $\text{plim}_{n\to\infty}\sqrt{n}(a_n - b_n) = 0$.
- The asymptotic variance or covariance of a random sequence a_n is said to be "equal to σ_a^2/n" if $\text{plim}_{n\to\infty}a_n =: a_\infty$ exists and if $\sqrt{n}(a_n - a_\infty)$ converges in distribution to $\text{N}(0, \sigma_a^2)$ as $n \to \infty$. The asymptotic variance or covariance of a_n is then denoted by $\text{var}(a_n) = \sigma_a^2/n$.

First note that, by (7) and (17),

$$\tilde{b}_c = F(\tilde{\mathcal{S}}) \;, \tag{20}$$

where F is a continuously differentiable function of

$$\tilde{\mathcal{S}} := \begin{pmatrix} \text{vech}(\tilde{S}) \\ \tilde{s}_{xy} \\ \tilde{s}_{xh} \\ \tilde{s}_{yh} \\ \tilde{s}_{hh} \end{pmatrix} \;. \tag{21}$$

The vector vech($\tilde{S}$) contains the lower triangular elements of $\tilde{S}$. Denote the probability limit of $\tilde{\mathcal{S}}$, which is known from Lemma 1, by $\bar{\mathcal{S}}$. Thus

$$\bar{\mathcal{S}} = \begin{pmatrix} \text{vech}(\tilde{\Sigma}) \\ \tilde{\sigma}_{xy} \\ \sigma_{xh} \\ \sigma_{yh} \\ \sigma_{hh} \end{pmatrix} . \tag{22}$$

The idea is now to show that

$$\tilde{\mathcal{S}} - \bar{\mathcal{S}} \ \sim \ G(\mathcal{S}) + \Delta \ , \tag{23}$$

where G is a continuously differentiable function of the second moments

$$\mathcal{S} := \begin{pmatrix} \text{vech}(S) \\ s_{xy} \\ s_{xh} \\ s_{yh} \\ s_{hh} \end{pmatrix} \tag{24}$$

based on the non-aggregated data. As will be shown, the "error vector" Δ is a function of the δ_i's defined in (8) and (9). Moreover, it is independent of $\mathcal{S}$. Thus, by computing the covariance matrices of $\mathcal{S}$ and Δ and by using the delta method, the asymptotic covariance matrix of $\tilde{\mathcal{S}}$ can be derived from (23). From (20), by using the delta method once more, one can finally obtain the asymptotic covariance matrix of $\tilde{b}_c$.

To prove (23), we introduce the following fundamental lemma:

Lemma 2. *Assume $Y, X_1, \ldots, X_p$, and H to be jointly normally distributed. Consider the inverse regression models (8) and (9). Let the empirical variances and covariances of the non-aggregated and aggregated values of δ_i and δ_j, $i, j = y, 1, \ldots, p$, be denoted by $s_{\delta_i\delta_j}$ and $\tilde{s}_{\delta_i\delta_j}$, respectively (they are defined in a similar way as (2) and (3)). The following relations hold for $i, j = y, 1, \ldots, p$:*

a) $\tilde{s}_{ij} - \tilde{\sigma}_{ij} \sim \frac{1}{A}(s_{ij} - \sigma_{ij}) + (1 - \frac{1}{A})\left(\frac{s_{ih}s_{jh}}{s_{hh}} - \frac{\sigma_{ih}\sigma_{jh}}{\sigma_{hh}}\right) + (\tilde{s}_{\delta_i\delta_j} - \frac{1}{A}s_{\delta_i\delta_j})$,

b) $\tilde{s}_{ih} - \sigma_{ih} \sim s_{ih} - \sigma_{ih}$,

c) $\tilde{s}_{hh} - \sigma_{hh} \sim s_{hh} - \sigma_{hh}$.

Proof: As H is a sorting variable related to $Y, X_1, \ldots, X_p$ by the linear regression models (8) and (9), Lemma 2 directly follows from Lemma 2 in Schmid and Schneeweiss (2005).

Lemma 2 can now be used to define the elements of Δ: Let $S_{\delta,xx} := (\tilde{s}_{\delta_i\delta_j} - \frac{1}{A}s_{\delta_i\delta_j})_{i,j=1,\ldots,p}$ and $S_{\delta,xy} := (\tilde{s}_{\delta_i\delta_y} - \frac{1}{A}s_{\delta_i\delta_y})_{i=1,\ldots,p}$. Then

$$\Delta := \begin{pmatrix} \mathrm{vech}(S_{\delta,xx}) \\ S_{\delta,xy} \\ \mathbf{0} \end{pmatrix} , \tag{25}$$

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 easily seen that equation (23) 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 $\tilde{s}_{\delta_i\delta_j} - \frac{1}{A}s_{\delta_i\delta_j}$. Moreover, it can be shown that $G(\mathcal{S})$ and Δ are asymptotically independent.

Next, we have to compute the asymptotic covariance matrix of Δ. To this purpose, we introduce another lemma:

Lemma 3. *For any $i, j = y, 1, \ldots, p$, the expressions $\Delta_{ij} := (\tilde{s}_{\delta_i\delta_j} - s_{\delta_i\delta_j}/A)$ are asymptotically jointly normally distributed with zero mean. The asymptotic covariance of Δ_{ij} and Δ_{mn}, $i, j, m, n = y, 1, \ldots, p$, is given by*

$$\sigma_{\Delta_{ij}\Delta_{mn}} := \frac{1}{n}\frac{A-1}{A^2}\left(\sigma_{\delta_i\delta_m}\sigma_{\delta_j\delta_n} + \sigma_{\delta_i\delta_n}\sigma_{\delta_j\delta_m}\right) . \tag{26}$$

Proof: Lemma 3 follows from the Lemma 3 in Schmid and Schneeweiss (2005).

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

Now, by applying the delta method, we obtain

$$\mathrm{cov}(\tilde{\mathcal{S}}) = D_G\,\mathrm{cov}(\mathcal{S})D_G' + \Sigma_\Delta , \tag{27}$$

where D_G is the Jacobian matrix of $G(\mathcal{S})$ evaluated at $\mathrm{plim}_{n\to\infty}\mathcal{S}$.

The covariance matrix of $\mathcal{S}$ in (27) 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 $S_{Y,X,H}$ follows a Wishart$(p+2, n-1, \Sigma_{Y,X,H})$ distribution, we have

$$\operatorname{cov}(s_{ij}, s_{mn}) \;=\; \frac{1}{n}\left(\sigma_{im}\sigma_{jn} + \sigma_{in}\sigma_{jm}\right) \;,\;\; i, j, m, n = y, 1, \ldots, p, h \tag{28}$$

(compare Evans *et al.* (1993), p. 158).

From (27), by applying the delta method once more, we finally obtain

$$\operatorname{var}(\tilde{b}_c) \;=\; D_F \left(D_G \operatorname{cov}(\mathcal{S}) D_G' + \Sigma_\Delta\right) D_F' \;, \tag{29}$$

where D_F is the Jacobian matrix of $F(\tilde{\mathcal{S}})$ evaluated at $\bar{\mathcal{S}}$. Obviously, as seen from (26) and (28), $\operatorname{var}(\tilde{b}_c)$ is a function of the variances and covariances $\sigma_{\delta_i\delta_j}$ and σ_{ij} and also of $\tilde{\Sigma}$ and $\tilde{\sigma}_{xy}$. The asymptotic variance of $\tilde{b}_c$ can thus be estimated by replacing

- σ_{ih}, $i = y, 1, \ldots, p$, with their consistent estimators $\tilde{s}_{ih}$, $i = y, 1, \ldots, p$,
- σ_{hh} with its consistent estimator $\tilde{s}_{hh}$,
- $\sigma_{\delta_i\delta_j}$, $i, j = y, 1, \ldots, p$, with their consistent estimators (see Corollary 1 in Schmid and Schneeweiss (2005))

$$\tilde{\sigma}_{\delta_i\delta_j,c} \;:=\; A\left(\tilde{s}_{ij} - \frac{\tilde{s}_{ih}\tilde{s}_{jh}}{\tilde{s}_{hh}}\right) \;,\;\; i, j = y, 1, \ldots, p, \tag{30}$$

- σ_{ij}, $i, j = y, 1, \ldots, p$, with their consistent estimators (see equation (33) in Schmid and Schneeweiss (2005))

$$\tilde{\sigma}_{ij,c} \;:=\; A\tilde{s}_{ij} + (1 - A)\frac{\tilde{s}_{ih}\tilde{s}_{jh}}{\tilde{s}_{hh}} \;,\; i, j = y, 1, \ldots, p, \tag{31}$$

- $\tilde{\Sigma}$ with $\tilde{S}$,
- $\tilde{\sigma}_{xy}$ with $\tilde{s}_{xy}$.

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 purpose, we carried out a systematic simulation study using the statistical software R. The model we studied was a linear regression with two normally distributed covariates X_1 and X_2. The variance parameters have been chosen to be $\sigma_{11} = 1$, $\sigma_{22} = 4$, and $\sigma_{12} = 1$, which corresponds to a correlation of 0.5 between the two covariates.

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

To study the bias of $\tilde{b}_c$, we took $A = 3$ (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 σ_{11} and σ_{22}.

Now, for various values of β_1, the bias of $\tilde{b}$ and $\tilde{b}_c$ was estimated from 1000 randomly generated data sets (x_{z1}, x_{z2}, y_z), $z = 1, \ldots, n$. 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, 4. the regressor X_1. In Figs. 1 - 4, bias$(\tilde{b}_1)$ and bias$(\tilde{b}_{1,c})$ are plotted vs. β_1 for $n = 150$ and $n = 600$. Apparently, the finite sample bias of $\tilde{b}_{1,c}$ is close to zero if $n \geq 150$. Moreover, it can be seen from Figs. 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: In this case $\tilde{b}_1$ 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 $\tilde{b}_{1,c}$.

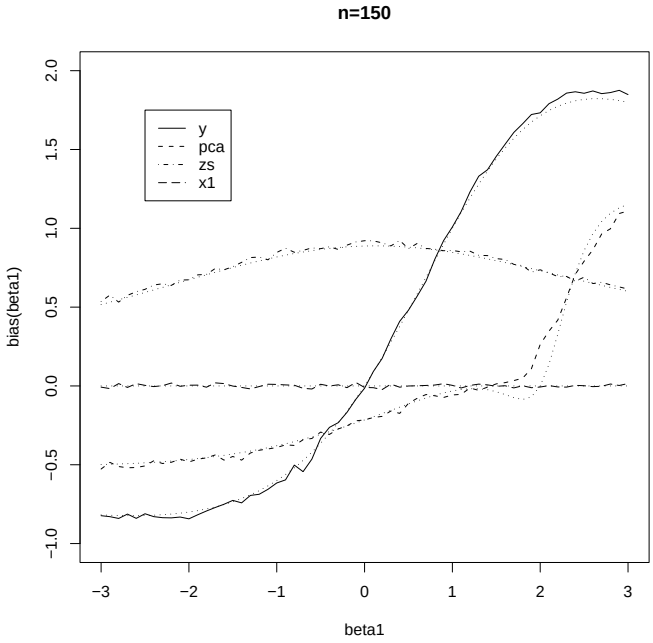

Figure 1: Bias of $\tilde{b}_1$ for various sorting variables (pca = first principal component projection, zs = sum of z-scores), dotted lines = true asymptotic bias curves

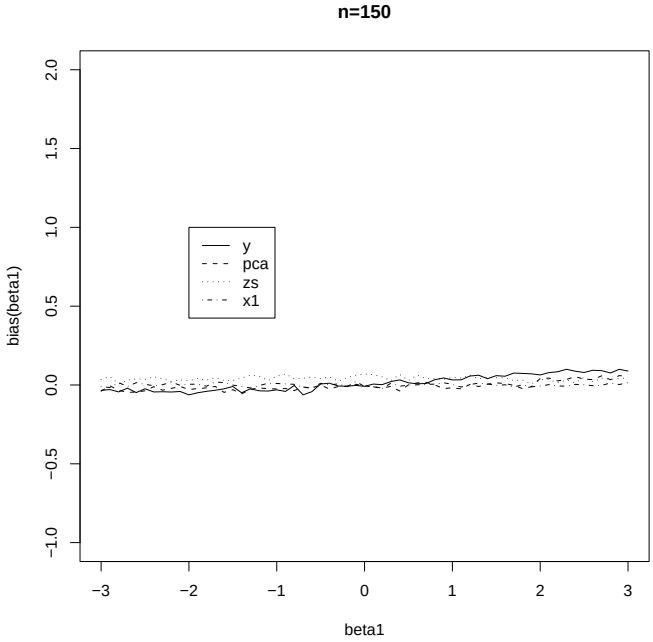

Figure 2: Bias of $\tilde{b}_{1,c}$ for various sorting variables (pca = first principal component projection, zs = sum of z-scores)

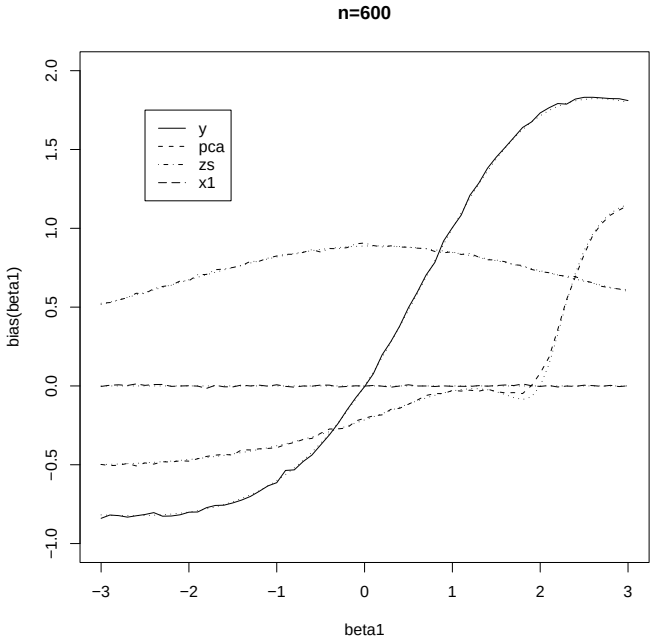

Figure 3: Bias of $\tilde{b}_1$ for various sorting variables (pca = first principal component projection, zs = sum of z-scores), dotted lines = true asymptotic bias curves

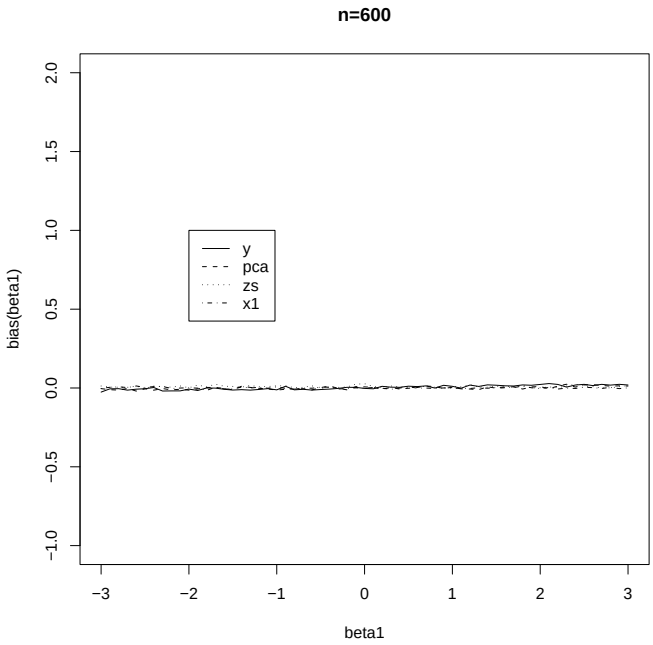

Figure 4: Bias of $\tilde{b}_{1,c}$ for various sorting variables (pca = first principal component projection, zs = sum of z-scores)

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

Figs. 5 and 6 contain the variances of $\sqrt{n}\,\tilde{b}_{1,c}$, which were estimated from the above simulation study for $n = 150$ and $n = 600$. Moreover, Figs. 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 ($n = 150$), $\text{var}(\tilde{b}_{1,c})$ is underestimated by its asymptotic counterpart. For large sample sizes ($n = 600$), we see that 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 $\tilde{b}_{1,c}$.

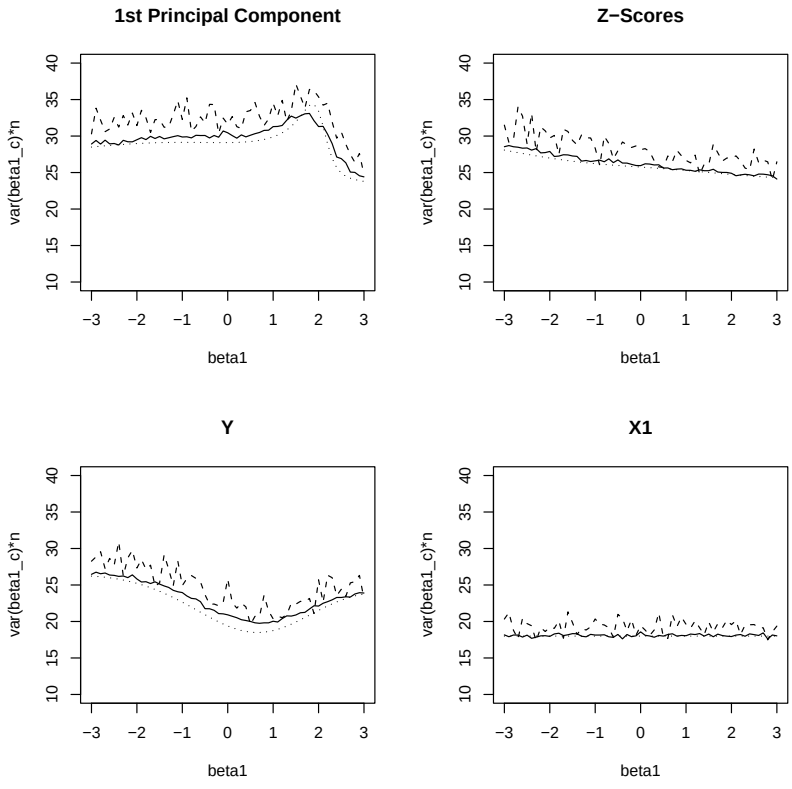

Figure 5: Variance of $\sqrt{n}\,\tilde{b}_{1,c}$, estimated from simulation (dashed lines: true variances, solid lines: averages of the estimated asymptotic variances, dotted lines: true asymptotic variances) - $n = 150$

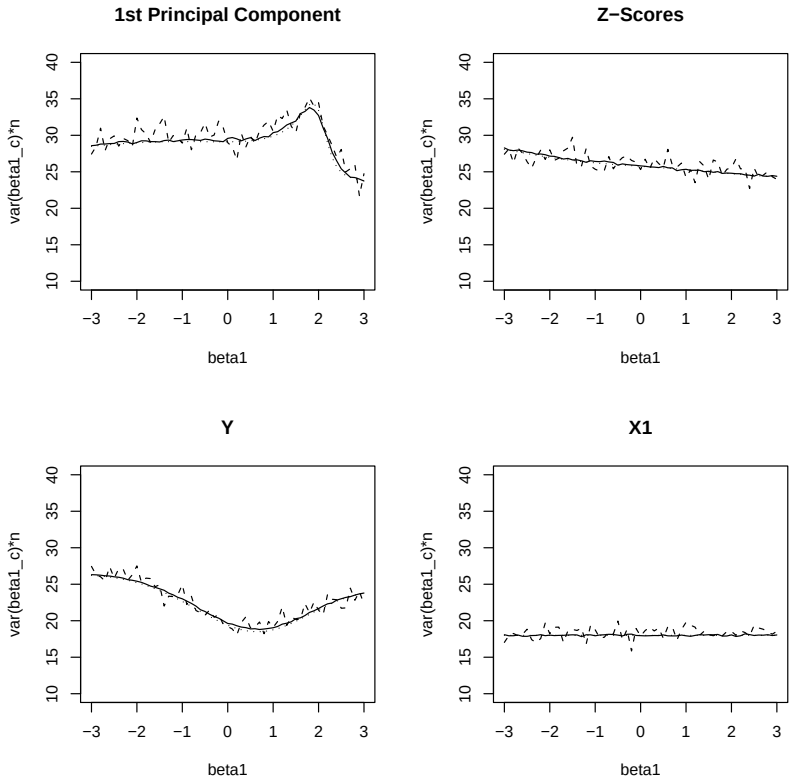

Figure 6: Variance of $\sqrt{n}\,\tilde{b}_{1,c}$, estimated from simulation (dashed lines: true variances, solid lines: averages of the estimated asymptotic variances, dotted lines: true asymptotic variances) - $n = 600$

6 Analysis of the Munich Rent Data

The simulation results presented in Section 5 are based on samples drawn from a multivariate normal distribution. In fact, the joint normality of the variables in model (1) is one of the key assumptions made to derive the asymptotic covariance matrix of $\tilde{b}_c$. In practice, however, the normality assumption will usually not hold.

In order to see how our method works in practice and also to find out how sensitive our results are with respect to deviations

from the normality assumption, we applied our estimation method to the 2003 Munich Rent Data (`http://www.statistik.lmu.de/service/datenarchiv/miete/miete03_e.html`), which certainly deviate from normality (see later). The data set contains 2053 households interviewed for the 2003 Munich rent standard. As it is publicly available, the *original* parameter estimates can be computed, and the impact of microaggregation on a linear regression can be studied directly. We are interested in the relationship between the monthly net rent of the households in EUR (`nr`, dependent variable), the floor space in m^2 (`fs`, independent variable), and the year of construction of the buildings (`yc`, independent variable). These variables clearly are not normally distributed (compare Fig. 7).

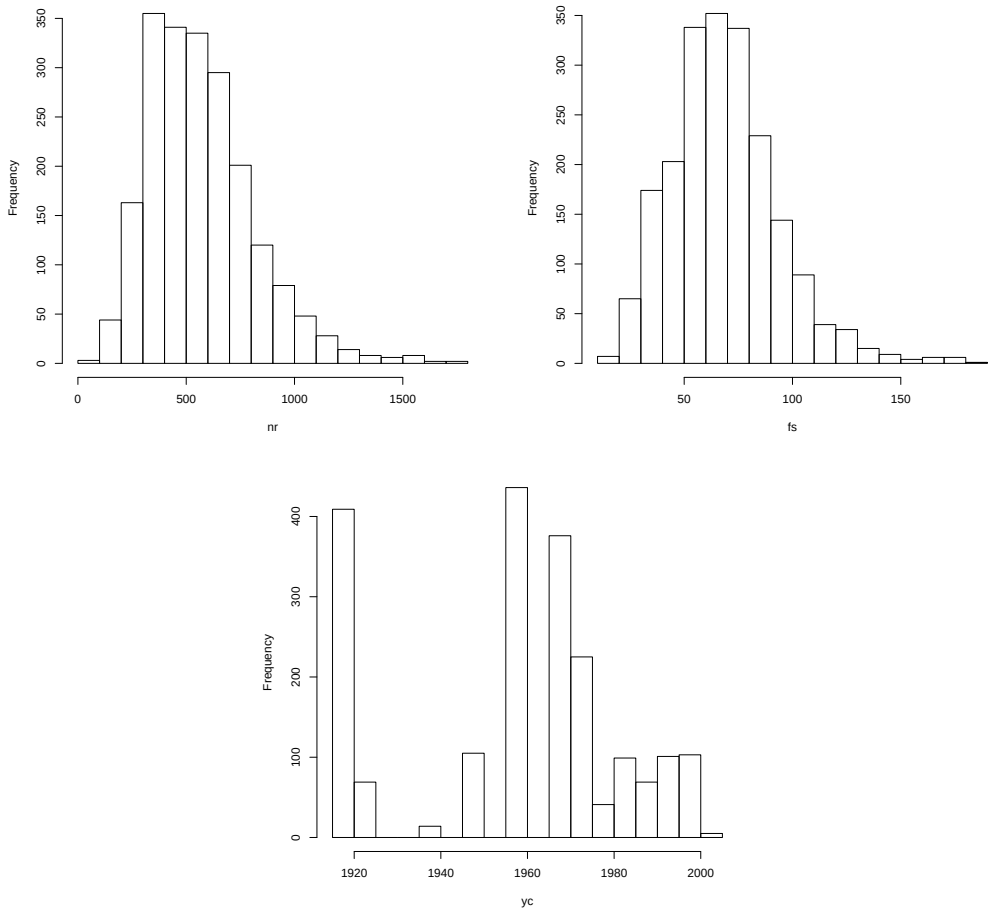

Figure 7: Histograms of `nr`, `fs`, and `yc`

To see whether our results hold despite the non-normality of `nr`, `fs`, and `yc`, we estimated a linear model based on the original (non-aggregated) data. We then compared the resulting estimates to the linear model estimates based on the microaggregated data set with group size $A = 3$. The sorting variables we used are the same as those used in Section 5. To obtain a sample size which is a multiple of A (i.e. $n = 2052$), we sorted the original data set with respect to `nr` and deleted the median observation.

The naive linear model estimates are shown in Table 1. As expected, microaggregation with respect to *pca*, *zs*, or the dependent variable `nr` leads to biased estimates of the model parameters β_{fs} and β_{yc}. In contrast to the results of Section 3, the naive estimators are also severely biased if one of the regressors (`fs` or `yc`) is used as the sorting variable. Although `nr`, `fs`, and `yc` are not normally distributed, this is a surprising result: Feige and Watts (1972) showed that in a linear model, $\tilde{b}$ is an unbiased estimator of β if H is independent of ϵ, no matter what distribution(s) the variables follow. Thus the results in rows 6 and 7 of Table 1 are probably due to the large variance of $\tilde{b}$. A more plausible result is obtained if a "z-score" linear combination of the regressors is used as the sorting variable: In this case, the naive estimators of β_{fs} and β_{yc} are remarkably close to the original estimates (see the last row in Table 1). It thus seems recommendable to use a linear combination of the regressors to sort the data, instead of one single regressor.

Table 2 shows that the correction of $\tilde{b}$ works as it should: The corrected estimates are close to the original estimates, implying that the corrected estimators are robust against violations of the normality assumption. We also see that the standard errors of the parameter estimates, as estimated by the procedure described in Section 4, are larger than the standard errors

Sorting variable	$\tilde{b}_{fs}$	$\tilde{b}_{yc}$	$\hat{\sigma}_{\tilde{b}_{fs}}$	$\hat{\sigma}_{\tilde{b}_{yc}}$
non-aggregated data	7.28	1.93	0.15	0.15
`nr`	10.20	2.56	0.13	0.19
pca	10.40	2.60	0.12	0.17
zs	8.78	2.64	0.11	0.13
`fs`	7.57	3.28	0.12	0.19
`yc`	9.90	2.47	0.13	0.11
$\texttt{fs}/\hat{\sigma}_{\texttt{fs}} + \texttt{yc}/\hat{\sigma}_{\texttt{yc}}$	7.39	1.83	0.10	0.11

Table 1: Regression of `nr` on `fs` and `yc` - naive estimates

Sorting variable	$\tilde{b}_{fs,c}$	$\tilde{b}_{yc,c}$	$\hat{\sigma}_{\tilde{b}_{fs,c}}$	$\hat{\sigma}_{\tilde{b}_{yc,c}}$	$\hat{\sigma}_{\tilde{b}_{fs,c,bs}}$	$\hat{\sigma}_{\tilde{b}_{yc,c,bs}}$
`nr`	6.82	1.71	0.21	0.22	0.24	0.19
pca	7.46	1.99	0.21	0.22	0.26	0.19
zs	7.36	1.68	0.19	0.22	0.27	0.24
`fs`	7.57	3.28	0.21	0.33	0.20	0.31
`yc`	9.90	2.47	0.23	0.19	0.30	0.18
$\texttt{fs}/\hat{\sigma}_{\texttt{fs}} + \texttt{yc}/\hat{\sigma}_{\texttt{yc}}$	7.39	1.83	0.18	0.18	0.22	0.20

Table 2: Regression of `nr` on `fs` and `yc` - corrected estimates (columns 2 - 5) and bootstrap variance estimates (columns 6 & 7)

of the estimates based on the non-aggregated data. This is not a surprising result, as the application of anonymization techniques to data sets always implies an efficiency loss in parameter estimation.

To see whether the standard errors in rows 4 and 5 of Table 2 are reliable estimates of the true standard errors of $\tilde{b}_{fs,c}$ and $\tilde{b}_{yc,c}$, we additionally estimated $\text{var}(\tilde{b}_{fs,c})$ and $\text{var}(\tilde{b}_{yc,c})$ from 10000 bootstrap samples of size $n = 2052$. The results obtained are shown in columns 6 and 7 of Table 2. Apparently, most of the bootstrap variance estimates are close to their counterparts based on the multivariate normal distribution. Thus the correction procedure proposed in Sections 3 and 4 seems to be robust against violations of the model assumptions.

7 Estimation of the coefficients $c_y, c_1, \ldots, c_p$

The computation of the corrected estimator $\tilde{b}_c$ requires the aggregated data values of the sorting variable H to be known to the data user. This either implies that data holders have to provide the aggregated data vector $\tilde{h}$ or that data users are able to estimate $\tilde{h}$ from the aggregated data vectors $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$. In this section, we discuss some special cases of sorting variables where it is possible to estimate $\tilde{h}$ from the aggregated data.

7.1 Microaggregation with respect to a Particular Variable in the Data Set

Consider the case where H is an exact linear combination of $Y, X_1, \ldots, X_p$: In this case the aggregated data values of H can be easily reconstructed from $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$:

$$\tilde{h} = c_y\tilde{y} + c_1\tilde{x}_1 + \cdots + c_p\tilde{x}_p \ . \tag{32}$$

This implies 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, it is sufficient to tell data users which variable has been used to sort the data. As in this case all of the coefficients are 0 except one coefficient (which "belongs" to 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.

7.2 Microaggregation with respect to the First Principal Component Projection

If H is an arbitrary linear combination of $Y, X_1, \ldots, X_p$ and no information on $c_y, c_1, \ldots, c_p$ is provided by the data holder, the coefficients $c_y, c_1, \ldots, c_p$ have to be estimated from the aggregated data. In the following, we will show that $c_y, c_1, \ldots, c_p$ can be consistently estimated if

1. H is the first principal component projection of $Y, X_1, \ldots, X_p$ and
2. the covariance matrix $\Sigma_{Y,X}$ of Y and $X_1, \ldots, X_p$ is used to compute the first principal component projection.

Lemma 4. *Let Σ_z be the covariance matrix of $Z := (Y, X_1, \ldots, X_p)$ and $\tilde{\Sigma}_z$ be the probability limit of the empirical covariance matrix $\tilde{S}_z$ based on the aggregated data. Denote the eigenvector associated to the largest eigenvalue λ of Σ_z by $c = (c_y, c_1, \ldots, c_p)'$. Then c is also the eigenvector corresponding to the largest eigenvalue of $\tilde{\Sigma}_z$.*

Proof. We first show that

$$\tilde{\Sigma}_z c = \Sigma_z c \ . \tag{33}$$

To prove equation (33) we make use of the relationship $H = Z'c$. Thus

$$\sigma_{hh} = c'\Sigma_z c \,, \tag{34}$$
$$\sigma_{zh} = \Sigma_z c \,, \tag{35}$$

where σ_{zh} is the covariance vector of Z and H. From (34), (35), Lemma 1d) and e), and from a corresponding formula for $\text{plim}_{n\to\infty}\tilde{s}_{yy}$, it follows that

$$\tilde{\Sigma}_z = \frac{1}{A}\Sigma_z + \Big(1 - \frac{1}{A}\Big)\frac{\Sigma_z c c'\Sigma_z}{c'\Sigma_z c} \,. \tag{36}$$

We thus obtain

$$\tilde{\Sigma}_z c = \frac{1}{A}\Sigma_z c + \Big(1 - \frac{1}{A}\Big)\Sigma_z c = \Sigma_z c = \lambda c \,. \tag{37}$$

Note that equation (33) not only holds for the eigenvector c corresponding to the largest eigenvalue of Σ_z but for *any* vector of coefficients c that is used to construct the aggregated values of H (given that H is an exact linear combination of $Y, X_1, \ldots, X_p$).

From (37) it follows that λ is an eigenvalue of $\tilde{\Sigma}_z$. To prove Lemma 4, we have to show that λ is also the *largest* eigenvalue of $\tilde{\Sigma}_z$.

As λ is defined as the largest eigenvalue of Σ_z, the following relationship holds:

$$\lambda = \max_d \frac{d'\Sigma_z d}{d'd} \,, \tag{38}$$

where d is an arbitrary vector of length $p+1$. Clearly, the right hand side of (38) is maximized by c. It thus remains to show that

$$\max_d \frac{d'\tilde{\Sigma}_z d}{d'd} = \frac{c'\tilde{\Sigma}_z c}{c'c} \;(= \lambda) \,. \tag{39}$$

Without loss of generality we only consider vectors $d \in \mathbb{R}^{p+1}$ with $\|d\| = d'd = 1$. Note that each d can be written as a linear combination of the (orthonormal) eigenvectors $c, p_1, \ldots, p_p$ of Σ_z:

$$d = uc + \sum_{i=1}^{p} u_i p_i \,, \tag{40}$$

where $|u| \leq 1$, $|u_i| \leq 1$, $i = 1, \ldots, p$. It follows that

$$\begin{aligned}
\frac{d'\tilde{\Sigma}_z d}{d'd} &= \frac{1}{A}\frac{d'\Sigma_z d}{d'd} + \Big(1 - \frac{1}{A}\Big)\frac{(d'\Sigma_z c)(c'\Sigma_z d)}{d'd \cdot c'\Sigma_z c} \\
&= \frac{1}{A} d'\Sigma_z d + \Big(1 - \frac{1}{A}\Big)\frac{(c'\Sigma_z d)^2}{c'\Sigma_z c} \\
&= \frac{1}{A} d'\Sigma_z d + \Big(1 - \frac{1}{A}\Big)\frac{(\lambda c'(uc + \sum_{i=1}^{p} u_i p_i))^2}{c'\Sigma_z c} \\
&= \frac{1}{A} d'\Sigma_z d + \Big(1 - \frac{1}{A}\Big)\frac{\lambda^2 u^2}{c'\Sigma_z c} \qquad (41)
\end{aligned}$$

Due to (38), the left summand in (41) is maximized by c. As $|u| \leq 1$, the right summand in (41) is maximal if $u = \pm 1$ (implying that $d = \pm c$). This proves (39).

□

Lemma 4 implies that c can be consistently estimated by the eigenvector associated with the largest eigenvalue of the empirical covariance matrix $\tilde{S}_z$ based on the aggregated data. Thus the corrected estimator $\tilde{b}_c$ (using the reconstructed aggregated data values of H) will be a consistent estimator of β. While the estimation of c introduces additional variance to the corrected estimator $\tilde{b}_c$, simulations similar to those performed in Section 5 show that the increase in variance is negligible.

In practice, however, it is more common to use the correlation matrix of $Y, X_1, \ldots, X_p$ instead of the covariance matrix Σ_z to determine the first principal component projection. Note that in this case, the results of Lemma 4 cannot be applied, and another estimation technique has to be applied to estimate the coefficients $c_y, c_1, \ldots, c_p$ (see Subsection 7.3).

7.3 Microaggregation with respect to the Sum of Z-Scores

If H is the sum of z-scores, it is convenient to consider the system of equations

$$A\tilde{\Sigma}_z - (A-1)\frac{\Sigma_z c c' \Sigma_z}{c'\Sigma_z c} - \Sigma_z = 0 \,, \tag{42}$$

$$\left(\frac{1}{\sqrt{\sigma_{yy}}}, \frac{1}{\sqrt{\sigma_{11}}}, \ldots, \frac{1}{\sqrt{\sigma_{pp}}}\right)' = c \,, \tag{43}$$

where (42) is just a reformulation of (36). By using the empirical covariance matrix $\tilde{S}_z$ instead of $\tilde{\Sigma}_z$ and by solving (42) and (43) numerically, we can obtain an estimate $\hat{\Sigma}_z$ of the covariance matrix Σ_z. Thus, an estimate of c is given by

$$\hat{c} := \left(\frac{1}{\sqrt{\hat{\sigma}_{yy}}}, \frac{1}{\sqrt{\hat{\sigma}_{11}}}, \ldots, \frac{1}{\sqrt{\hat{\sigma}_{pp}}}\right)' \,, \tag{44}$$

where $\hat{\sigma}_{yy}, \hat{\sigma}_{11}, \ldots, \hat{\sigma}_{pp}$ are the diagonal elements of $\hat{\Sigma}_z$. The estimates of the coefficients $c_y, c_1, \ldots, c_p$ can then be used to construct $\tilde{h}$, and the correction of $\tilde{\beta}$ can be performed as described in Sections 3 and 4.

Note that the above described procedure to estimate the vector of coefficients c can be applied for any c that depends on the second order moments of $Y, X_1, \ldots, X_p$. For example, if H is the first principal component projection based on the correlation matrix of $Y, X_1, \ldots, X_p$, the system of equations

$$A\tilde{\Sigma}_z - (A-1)\frac{\Sigma_z c c' \Sigma_z}{c'\Sigma_z c} - \Sigma_z = 0 \,, \tag{45}$$

$$\operatorname{diag}(\Sigma_z)^{-1/2}\,\Sigma_z\,\operatorname{diag}(\Sigma_z)^{-1/2} c = \lambda c \tag{46}$$

has to be solved instead of (42) and (43). Again, simulations similar to those performed in Section 5 show that the increase in variance induced by the additional estimation of c is small if compared to the variance increase induced by microaggregation.

8 Conclusion

We have analyzed the effects of single-axis-sorting microaggregation on the estimation of a linear regression model in continuous variables. In Section 3 we have shown that the naive least squares estimators of a linear model are not necessarily consistent estimators of the true model parameters. However, in the special case where the sorting variable is a linear combination of the regressors, the naive 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 more convenient for statistical analysis, it has to be pointed out that data providers 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 is a very relevant case.

The main result of the paper is the development of a consistent estimator that removes the aggregation bias of the naive LS estimator of β. We also derived the asymptotic covariance matrix of the corrected estimator for the slope parameter vector β. The simulation study in Section 5 as well as the analysis of the Munich Rent Data in Section 6 show that the correction procedure already works well if the sample size is moderately high ($n \geq 150$).

To prove our results, we assumed the variables in the linear model and the sorting variable to be jointly normally distributed. Although this assumption usually does not hold in practice, the analysis of the Munich Rent Data shows that the estimation procedure is robust against deviations from normality. The only exception is the case where a single regressor is used as the sorting variable: In this case the naive estimates (which are equal to the corrected estimates) turn out to be severely biased.

In order 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. If the sorting variable is an exact linear combination of the variables in the linear model, data holders only have to provide the coefficients of this combination. Moreover, there are special types of sorting variables where the coefficients can be consistently estimated from the aggregated data set, see Section 7. Simulations show that the additional variance induced by the estimation of the coefficients is negligible in most cases.

In summary, we have derived a method that allows data users to consistently estimate linear models with microaggregated data. Thus the disclosure risk of data sets is kept low, while the results obtained from the anonymized data are still guaranteed to be analytically valid.

Acknowledgements

I thank Hans Schneeweiss for very helpful discussions and comments. Financial support from the Deutsche Forschungsgemeinschaft (German Science Foundation) is gratefully acknowledged.

References

Defays, D. and P. Nanopoulos (1993): "Panels of Enterprises and Confidentiality: The Small Aggregates Method," 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. and J. M. Mateo-Sanz (2002): "Practical Data-Oriented Microaggregation for Statistical Disclosure Control," IEEE Transactions on Knowledge and Data Engineering, 14, No. 1, 189-201.

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

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

Schmid, M. and H. Schneeweiss (2005): "Estimation of a Linear Regression under Microaggregation with the Response Variable as a Sorting Variable," Discussion Paper 462, SFB 386, Department of Statistics, University of Munich.

Statistisches Bundesamt (2005): *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten.* Statistik und Wissenschaft 4, Wiesbaden: Statistisches Bundesamt.