Schmid, Schneeweiß:

Estimation of a Linear Regression under Microaggregation with the Response Variable as a Sorting Variable

Sonderforschungsbereich 386, Paper 462 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Estimation of a Linear Regression under Microaggregation with the Response Variable as a Sorting Variable

Matthias Schmid and Hans Schneeweiss

Department of Statistics, University of Munich
Ludwigstr. 33, 80539 Munich, Germany

Abstract

Microaggregation is one of the most frequently applied 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. However, while reducing the disclosure risk of data files, the technique also affects the results of statistical analyses. The paper deals with the impact of microaggregation on a linear model in continuous variables. We show that parameter estimates are biased if the dependent variable is used to form the groups. 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.

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 (using a 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.

The main problem with microaggregation is that traditional statistical estimation techniques may be severely biased and less efficient if applied to the microaggregated data. Thus, in order to reduce the information loss imposed by microaggregation, only those data values which are "similar" to each other should be grouped (see Feige and Watts (1972)). In the literature, many suggestions have been made on how to best form the groups (Domingo-Ferrer and Mateo-Sanz (2002), Laszlo and Mukherjee (2005)). However, the impact of these techniques on parameter estimation, hyptothesis tests, etc., has still to be investigated.

The present paper deals with microaggregation by a *sorting variable*, one of the oldest and most frequently applied microaggregation techniques (Paass and Wauschkuhn (1985), Mateo-Sanz and Domingo-Ferrer (1998)). This procedure uses a fixed group size. The sorting variable can either be one of the regressors or the dependent variable in a statistical model. Groups are then formed by observations having similar values for the sorting variable.

Our aim is to investigate the effects of this technique on the least squares (LS) estimation of a linear regression in continuous variables. While the naive LS estimator remains unbiased if one of the covariates is used as the sorting

variable (see Feige and Watts (1972)), an extensive simulation study performed by Schmid and Schneeweiss (2005) has shown that microaggregation induces a severe bias if the dependent variable is used as the sorting variable. Although aggregating with respect to a covariate 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 variable will serve as the dependent one. Thus, investigating microaggregation with respect to the dependent variable is a very relevant case.

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 the dependent variable. 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.

The paper generalizes previous results for the *simple* linear regression (see Schmid *et al.* (2005a,b)) to the case of of a *multiple* linear regression. In addition to the arguments of the previous papers, some new lemmas are needed to prove the results of this paper.

Section 2 starts with a brief description of microaggregation by a sorting variable. 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. The results of this paper are summarized in Section 7. Proofs of lemmas are given in the appendix.

2 Microaggregation by a Sorting Variable

We consider microaggregation with respect to a sorting variable in the data set. In a linear model, the sorting variable can either be the dependent variable or one of the covariates. 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 linear model with two covariates X_1 and X_2 and a dependent variable Y. Assume the data set to be

x_1	2	1	4	7	3	4
x_2	1	3	4	2	8	6
y	2	7	6	8	3	1

.

Now, if Y is the sorting variable and $A = 3$, we obtain the sorted data set

$x_{1,sort}$	4	2	3	4	1	7
$x_{2,sort}$	6	1	8	4	3	2
y_{sort}	1	2	3	6	7	8

and the microaggregated data set

$\tilde{x}_1$	3	3	3	4	4	4
$\tilde{x}_2$	5	5	5	3	3	3
$\tilde{y}$	2	2	2	7	7	7

.

3 Consistent Estimation

3.1 Notation

As microaggregation with respect to a covariate leads to unbiased linear model estimates (compare Feige and Watts (1972)), we only consider microaggregation with respect to the dependent variable. In the following, the effect of this type of microaggregation on the LS estimation of the multiple linear regression model

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

is investigated. Y denotes the response (or endogenous) variable, while $X_1, \ldots, X_p$ denote the covariates (or exogenous variables). Y and $X_1, \ldots, X_p$ are assumed to be continuous random variables with variances σ_{yy}, $\sigma_{11}, \ldots, \sigma_{pp}$. The supports of $Y, X_1, \ldots, X_p$ are (possibly infinite) intervals.

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. The vectors containing the aggregated data are denoted by $\tilde{y}$ and $\tilde{x}_1, \ldots, \tilde{x}_p$. For simplicity, it is assumed throughout that n is a multiple of A. Note that in this case, the empirical means $\bar{y}, \bar{x}_1, \ldots, \bar{x}_p$ of $y, x_1, \ldots, x_p$ are the same as the empirical means $\bar{\tilde{y}}, \bar{\tilde{x}}_1, \ldots, \bar{\tilde{x}}_p$ of $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$, 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}, $i = 1, \ldots, p$, and the variance of Y by σ_{yy}.

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 $\sigma := (\sigma_{iy})_{i=1,\ldots,p}$ be the covariance (column) vector of $(X_1, \ldots, X_p)$ and Y. The empirical variances of y and $\tilde{y}$ are denoted by s_{yy} and $\tilde{s}_{yy}$, respectively, and the empirical covariances of x_i and y and of $\tilde{x}_i$ and $\tilde{y}$ are denoted by s_{iy} and $\tilde{s}_{iy}$, respectively. Finally let

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

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 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

$$\tilde{b} := \tilde{S}^{-1}\tilde{s} \ . \tag{5}$$

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

Lemma 1. *Assume that there exist inverse linear relationships*

$$X_i = \alpha_i + \gamma_i Y + \delta_i \ , \ i = 1, \ldots, p \ , \tag{6}$$

where the δ_i's are random variables, independent of Y, with zero mean and variances and covariances $\sigma_{\delta_i\delta_j}$, $1 \leq i, j \leq p$. Then the following probability limits exist:

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

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

c) $\text{plim}_{n\to\infty}\tilde{S} = \frac{1}{A}\Sigma + \left(1 - \frac{1}{A}\right)\frac{\sigma\sigma'}{\sigma_{yy}} =: \tilde{\Sigma}$.

Proof: See appendix.

With Lemma 1, the probability limit of $\tilde{b}$ can be evaluated as

$$\begin{aligned} \tilde{\beta} &:= \text{plim}_{n\to\infty}\tilde{b} = \tilde{\Sigma}^{-1}\sigma = A\Big(\Sigma + \frac{A-1}{\sigma_{yy}}\sigma\sigma'\Big)^{-1}\sigma \\ &= A\left(\Sigma^{-1} - \Big(1 + \frac{A-1}{\sigma_{yy}}\sigma'\Sigma^{-1}\sigma\Big)^{-1}\frac{A-1}{\sigma_{yy}}\Sigma^{-1}\sigma\sigma'\Sigma^{-1}\right)\sigma\ . \end{aligned} \tag{7}$$

In order to obtain (7), we used a matrix inversion formula which can be found, e.g., in Dhrymes (1984), Corollary 5. With some algebra and using $\beta = \Sigma^{-1}\sigma$, it follows that

$$\tilde{\beta} = \frac{A}{1 + (A-1)\sigma'\Sigma^{-1}\sigma/\sigma_{yy}}\,\beta\ . \tag{8}$$

Thus the asymptotic bias of $\tilde{b}$ depends on the multiple correlation coefficient $R^2 := \sigma'\Sigma^{-1}\sigma/\sigma_{yy}$. This coefficient is always smaller than or equal to 1, so that β is asymptotically *overestimated* by the naive LS estimator $\tilde{b}$. The only exceptions are the following two cases:

1. $R^2 = 1$ (which corresponds to a perfect linear relationship between Y and $X_1, \ldots, X_p$). In this case $\tilde{\beta} = \beta$.
2. $R^2 = 0$ (in which case $\beta = 0$ and thus also $\tilde{\beta} = 0$).

In a simple linear model with one covariate X_1, equation (8) reduces to

$$\tilde{\beta} = \frac{A}{1 + (A-1)\rho^2}\,\beta\ , \tag{9}$$

where ρ is the correlation between Y and X_1. This is the same relationship as the one derived in Schmid *et al.* (2005a), Section 4.

From (8), we also see that the asymptotic bias of $\tilde{b}$ grows if the group size A becomes larger. As expected, for the non-aggregated data ($A = 1$), the bias factor in (8) is equal to 1.

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

$$\Sigma \;=\; \left(A\tilde{\Sigma} - (A-1)\,\frac{\sigma\sigma'}{\sigma_{yy}} \right) \tag{10}$$

from Lemma 1c). By algebraic manipulations similar to those that led to (8), this yields

$$\beta \;=\; \frac{1}{A - (A-1)\sigma'\tilde{\Sigma}^{-1}\sigma/\sigma_{yy}}\,\tilde{\beta}\;, \tag{11}$$

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

$$\tilde{b}_c \;:=\; \frac{1}{A - (A-1)\,\tilde{s}'\tilde{S}^{-1}\tilde{s}/\tilde{s}_{yy}}\,\tilde{b} \;=\; \frac{1}{A-(A-1)\tilde{R}^2}\,\tilde{b}\;, \tag{12}$$

where $\tilde{R}^2$ denotes the empirical multiple correlation coefficient based on the aggregated data. Note that the factor in front of $\tilde{b}$ is always positive and is less than 1 for $A > 1$ and $\tilde{R}^2 < 1$.

A consistent estimator of the intercept β_0 is simply given by

$$\tilde{b}_{0c} \;:=\; \bar{\tilde{y}} - (\tilde{b}_{1c}\bar{\tilde{x}}_1 + \cdots + \tilde{b}_{pc}\bar{\tilde{x}}_p)\;, \tag{13}$$

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

Furthermore, from (10) and (12), we obtain a consistent estimator of the residual variance $\sigma_\epsilon^2 = \sigma_{yy} - \beta'\Sigma\beta$:

$$\tilde{\sigma}^2_{\epsilon,c} \;:=\; \tilde{s}_{yy} - \tilde{b}_c' \left(A\tilde{S} - (A-1)\,\frac{\tilde{s}\tilde{s}'}{\tilde{s}_{yy}} \right) \tilde{b}_c\;. \tag{14}$$

With some algebra, we obtain

$$\tilde{\sigma}^2_{\epsilon,c} \;=\; \frac{A}{A-(A-1)\tilde{R}^2}\,\tilde{s}_{yy}\left(1-\tilde{R}^2\right) \;=\; \frac{A}{A-(A-1)\tilde{R}^2}\,\tilde{\sigma}^2_{\epsilon}\;, \qquad (15)$$

where $\tilde{\sigma}^2_{\epsilon}$ is the naive estimator of σ^2_{ϵ} based on the aggregated data. We thus see that σ^2_{ϵ} is systematically underestimated by $\tilde{\sigma}^2_{\epsilon}$.

4 Asymptotic Covariance of $\tilde{b}_C$

To derive the asymptotic covariance matrix of $\tilde{b}_c$, we need stronger assumptions than in Section 3: $Y, X_1, \ldots, X_n$ are now assumed to be jointly *normally* distributed random variables.

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^2_a/n" if $\text{plim}_{n\to\infty} a_n =: a_\infty$ exists and if $\sqrt{n}(a_n - a_\infty)$ converges in distribution to $\text{N}(0, \sigma^2_a)$ as $n \to \infty$. The asymptotic variance or covariance of a_n is then denoted by $\text{var}(a_n) = \sigma^2_a/n$.

First note that, by (5) and (12),

$$\tilde{b}_c = F(\tilde{\mathcal{S}})\;, \qquad (16)$$

where F is a continuously differentiable function of

$$\tilde{\mathcal{S}} := \begin{pmatrix} \text{vech}(\tilde{S}) \\ \tilde{s} \\ \tilde{s}_{yy} \end{pmatrix} . \tag{17}$$

The vector $\text{vech}(\tilde{S})$ 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

$$\bar{\mathcal{S}} = \begin{pmatrix} \text{vech}(\tilde{\Sigma}) \\ \sigma \\ \sigma_{yy} \end{pmatrix} . \tag{18}$$

The idea is now to show that

$$\tilde{\mathcal{S}} - \bar{\mathcal{S}} \ \sim \ G(\mathcal{S}) + \Delta \ , \tag{19}$$

where G is a continuously differentiable function of the second moments

$$\mathcal{S} := \begin{pmatrix} \text{vech}(S) \\ s \\ s_{yy} \end{pmatrix} \tag{20}$$

based on the non-aggregated data. As will be shown, the "error vector" Δ is a function of the δ_i defined in (6). 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 (19). From (16), by using the delta method once more, one can finally obtain the asymptotic covariance matrix of $\tilde{b}_c$.

To prove (19), we introduce the following fundamental lemma:

Lemma 2. *Assume $Y, X_1, \ldots, X_p$ to be jointly normally distributed. Consider the inverse regression models (6). Let the empirical variances and covariances of the non-aggregated and aggregated values of δ_i and δ_j, $1 \leq i, j \leq p$, 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 $i, j = 1, \ldots, p$:*

a) $\tilde{s}_{ij} - \tilde{\sigma}_{ij} \sim \frac{1}{A}(s_{ij} - \sigma_{ij}) + (1 - \frac{1}{A})\left(\frac{s_{iy}s_{jy}}{s_{yy}} - \frac{\sigma_{iy}\sigma_{jy}}{\sigma_{yy}}\right) + (\tilde{s}_{\delta_i\delta_j} - \frac{1}{A}s_{\delta_i\delta_j})$.

b) $\tilde{s}_{iy} - \sigma_{iy} \sim s_{iy} - \sigma_{iy}$.

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

Proof: See appendix.

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

$$\Delta := \begin{pmatrix} \text{vech}(S_\delta) \\ \mathbf{0} \end{pmatrix}, \tag{21}$$

where $\mathbf{0}$ is a $(p+1)$-dimensional vector of zeros. From Lemma 2 and from the definition of the elements of Δ, it is easily seen that equation (19) 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 $1 \leq i, j \leq p$, 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}, $1 \leq i, j, m, n \leq p$, is given by*

$$\sigma_{\Delta_{ij}\Delta_{mn}} := \frac{1}{n}\frac{A-1}{A^2}\left(\sigma_{\delta_i\delta_m}\sigma_{\delta_j\delta_n} + \sigma_{\delta_i\delta_n}\sigma_{\delta_j\delta_m}\right) . \tag{22}$$

Proof: See appendix.

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

$$\operatorname{cov}(\tilde{\mathcal{S}}) = D_G \operatorname{cov}(\mathcal{S}) D_G' + \Sigma_\Delta , \tag{23}$$

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

The covariance matrix of $\mathcal{S}$ in (23) can be derived as follows: Denote the covariance matrix of $(Y, X_1, \ldots, X_p)$ by $\Sigma_{Y,X}$ and the empirical covariance matrix of $(Y, X_1, \ldots, X_p)$ by $S_{Y,X}$. Now, as $S_{Y,X}$ follows a Wishart$(p+1, n-1, \Sigma_{Y,X})$ distribution, we have

$$\operatorname{cov}(s_{ij}, s_{mn}) = \frac{1}{n}(\sigma_{im}\sigma_{jn} + \sigma_{in}\sigma_{jm}) , \quad i, j, m, n = y, 1, \ldots, p \tag{24}$$

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

From (23), by applying the delta method once more, we finally obtain

$$\operatorname{var}(\tilde{b}_c) = D_F \left(D_G \operatorname{cov}(\mathcal{S}) D_G' + \Sigma_\Delta\right) D_F' , \tag{25}$$

where D_F is the Jacobian matrix of $F(\tilde{\mathcal{S}})$ evaluated at $\bar{\mathcal{S}}$. Obviously, as seen from (22) and (24), $\mathrm{var}(\tilde{b}_c)$ is a function of the variances and covariances $\sigma_{\delta_i\delta_j}$ and σ_{ij} and also of the covariance matrix $\tilde{\Sigma}$. The asymptotic variance of $\tilde{b}_c$ can thus be estimated by replacing

- σ_{iy}, $i = 1, \ldots, p$, with their consistent estimators $\tilde{s}_{iy}$, $i = 1, \ldots, p$,
- σ_{yy} with its consistent estimator $\tilde{s}_{yy}$,
- $\sigma_{\delta_i\delta_j}$, $i, j = 1, \ldots, p$, with their consistent estimators (see (30))

$$\tilde{\sigma}_{\delta_i\delta_j,c} := A\left(\tilde{s}_{ij} - \frac{\tilde{s}_{iy}\tilde{s}_{jy}}{\tilde{s}_{yy}}\right), \quad i, j = 1, \ldots, p, \tag{26}$$

- σ_{ij}, $i, j = 1, \ldots, p$, with their consistent estimators (see (33))

$$\tilde{\sigma}_{ij,c} := A\tilde{s}_{ij} + (1 - A)\frac{\tilde{s}_{iy}\tilde{s}_{jy}}{\tilde{s}_{yy}}, \quad i, j = 1, \ldots, p, \tag{27}$$

- $\tilde{\Sigma}$ with $\tilde{S}$.

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 carry out a systematic simulation study using the statistical software R. The model we study is 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 $\beta_0 = 1$. 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 $\tilde{b}_c$ was estimated from 1000 randomly generated data sets (x_{i1}, x_{i2}, y_i), $i = 1, \ldots, n$. In Figs. 1 and 2, bias$(\tilde{b})$ and bias$(\tilde{b}_c)$ are plotted vs. β_1 for various sample sizes. Apparently, the finite sample bias of $\tilde{b}_c$ is close to zero if $n \geq 150$. Moreover, it can be seen from Fig. 1 that the bias of $\tilde{b}$ does not converge to 0.

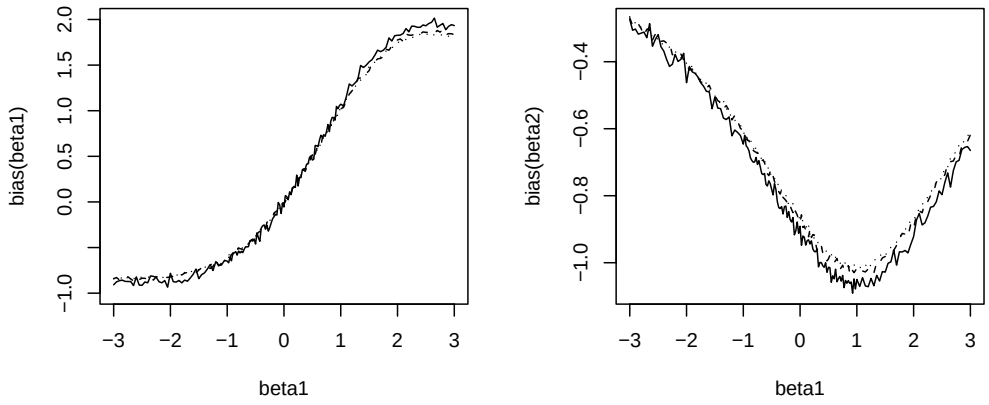

Figure 1: Bias of $\tilde{b}$ (solid line: $n = 51$, dashed line: $n = 150$, dotted line: $n = 300$)

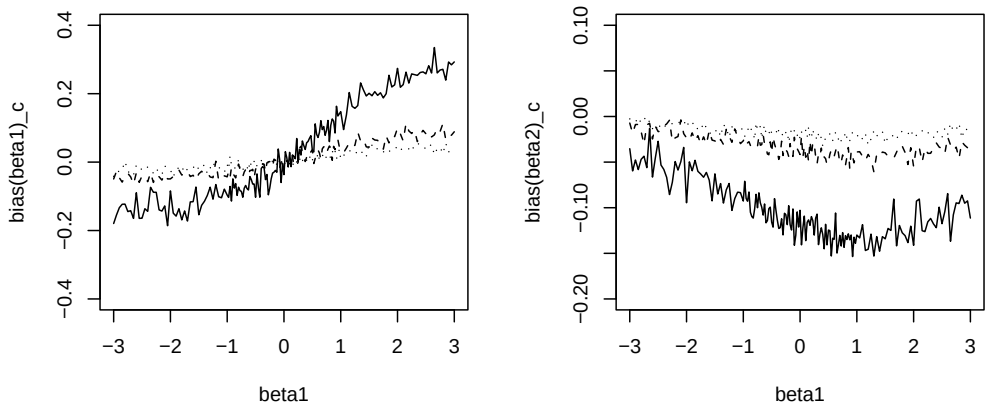

Figure 2: Bias of $\tilde{b}_c$ (solid line: $n = 51$, dashed line: $n = 150$, dotted line: $n = 300$)

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

Fig. 3 contains the variances and covariances of $\tilde{b}_{1,c}$ and $\tilde{b}_{2,c}$, which were estimated from the above simulation study. Moreover, Fig. 3 shows the averages of the estimated asymptotic variances and covariances of the elements of $\tilde{b}_c$, as well as the corresponding true asymptotic variances and covariances. We see that if the sample size is small ($n = 150$), $\text{var}(\tilde{b}_{1,c})$ and $\text{var}(\tilde{b}_{2,c})$ are underestimated by their asymptotic counterparts, whereas $\text{cov}(\tilde{b}_{1,c}, \tilde{b}_{2,c})$ is overestimated by its asymptotic counterpart. For large sample sizes ($n = 600$), we see that the asymptotic covariance matrix of $\tilde{b}_c$ is a good approximation of the true covariance matrix.

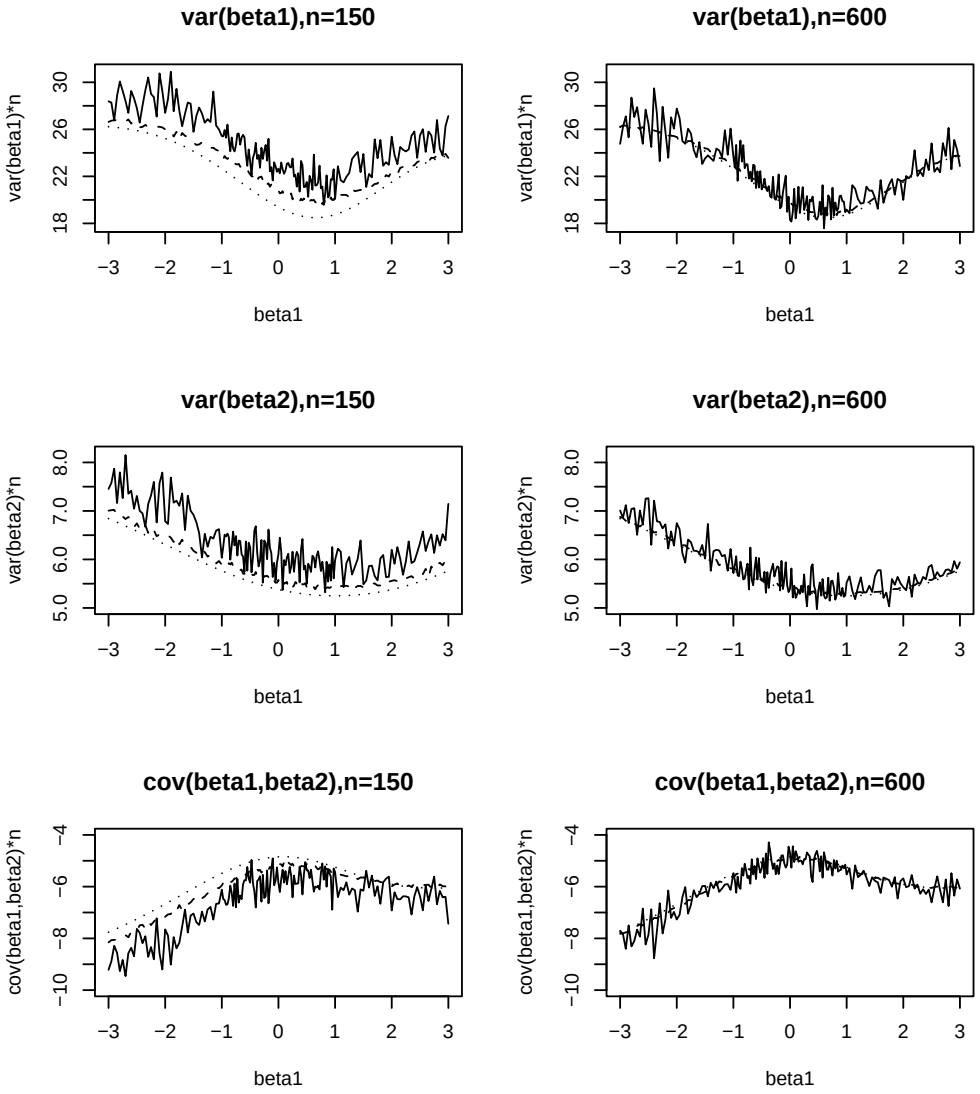

Figure 3: Variances and covariances of the elements of $\sqrt{n}\,\tilde{b}_c$, estimated from simulation (solid lines: true variances, dashed lines: averages of the estimated asymptotic variances, dotted lines: true asymptotic variances)

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. 4).

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 $A = 3$ and `nr` serving as

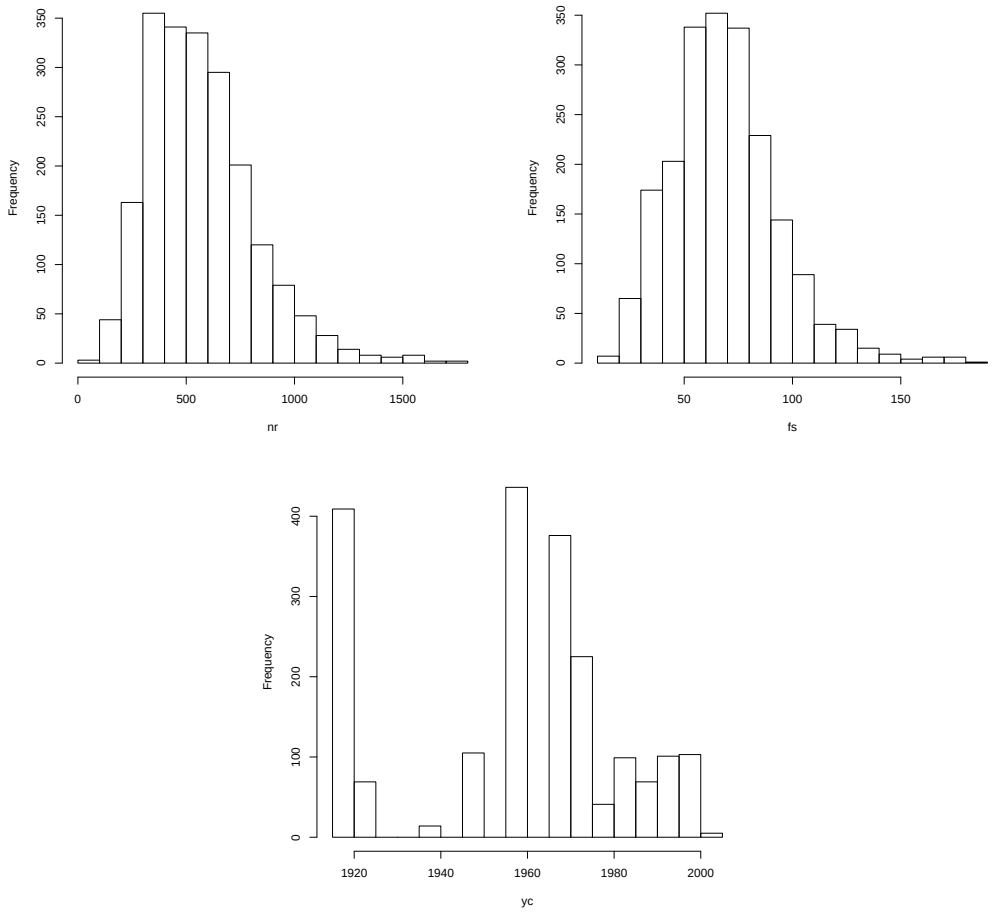

Figure 4: Histograms of `nr`, `fs`, and `yc`

the sorting variable. 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 linear model estimates are shown in Table 1. Row 1 contains the original parameter estimates based on the non-aggregated data. Row 2 contains the naive parameter estimates based on the microaggregated data, while the

	$\hat{\beta}_{fs}$	$\hat{\beta}_{yc}$	$\hat{\sigma}_{fs}$	$\hat{\sigma}_{yc}$	$\hat{\sigma}_{\epsilon}$
non-aggr. data	7.281	1.930	0.150	0.151	167.078
aggr. w.r.t. `nr` (naive)	10.201	2.556	0.130	0.191	122.175
aggr. w.r.t. `nr` (corr.)	6.824	1.710	0.212	0.224	172.990

Table 1: Regression of `nr` on `fs` and `yc`

corrected parameter estimates are contained in row 3. As expected, microaggregation with respect to `nr` leads to an overestimation of the effects of `fs` and `yc` by the naive LS estimators. Table 1 also shows that the correction of $\tilde{b}$ works as it should: The corrected estimates are close to the original estimates, although the standard errors of the parameter estimates, as estimated by the procedure described in Section 4, increase by about 45% (compared to the standard errors based on the non-aggregated data).

To see whether the standard errors in row 3 of Table 1 are reliable estimates of the true standard errors of $\tilde{b}_{fs,c}$ and $\tilde{b}_{yc,c}$, we additionally estimated $\mathrm{cov}(\tilde{b}_c)$ from 10000 bootstrap samples of size $n = 2052$. This procedure resulted in $\widehat{\mathrm{var}}(\tilde{b}_{fs,c}) = 0.243^2$ and $\widehat{\mathrm{var}}(\tilde{b}_{yc,c}) = 0.188^2$. Apparently, the bootstrap variance estimates are close to their counterparts based on the multivariate normal distribution (which take the values 0.212^2 and 0.224^2, respectively). We thus see that the correction procedure proposed in Sections 3 and 4 is robust against violations of the model assumptions.

7 Conclusion

We have analyzed the effects of microaggregation by a sorting variable on the estimation of a linear regression model in continuous variables. Feige and Watts (1972) have already shown that linear model estimates remain unbiased if one of the regressors is used to sort the data. We thus focused on the special case where the dependent variable is the sorting variable. We have shown that in this case, linear model estimates are asymptotically biased by a scalar factor. The bias factor is always greater than or equal to 1, which implies that the true slope parameters of the linear model are overestimated in absolute value. Moreover, the bias of the naive LS estimator depends on the multiple correlation coefficient R^2 of the dependent variable and the regressors. As $R^2 \to 1$, the asymptotic bias of the naive LS estimator tends to 0. In the special case where one of the slope parameters is equal to 0, the corresponding LS estimator of this parameter is asymptotically unbiased.

The main result of the present paper is the development of a consistent estimator that removes the aggregation bias of the naive LS estimator. 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 ($n \geq 300$).

We also derived the asymptotic covariance matrix of the corrected estimator for the slope parameter vector β. To do this, we assumed the dependent variable and the regressors to be jointly normally distributed. Although this assumption usually does not hold in practice, the analysis of the Munich Rent Data has shown that the estimation procedure is robust against deviations

from normality.

Future research includes the extension of the above results on all "single-axis sorting" microaggregation techniques. These techniques use an arbitrary linear combination of the dependent variable and the regressors to sort the data. For example, the sorting variable can be the first principle component projection or the sum of z-scores of the variables in a data set. The microaggregation technique considered in the present paper (where the dependent variable is the sorting variable) can thus be seen as a special case of single-axis sorting microaggregation. This implies that the correction procedure developed in this paper marks a starting point for a general evaluation of the bias induced by single-axis sorting microaggregation.

Appendix

Proof of Lemma 1: Part a) was proved in Schmid *et al.* (2005a). There it was also shown that $\text{plim}_{n\to\infty}\tilde{s}_{iy} = \sigma_{iy}$, from which part b) follows.

To derive the probability limit of $\tilde{s}_{ij}$, we make use of the relationships

$$\tilde{x}_i = \alpha_i + \gamma_i\tilde{y} + \tilde{\delta}_i \ , \ i = 1, \ldots, p \ , \tag{28}$$

where $\tilde{\delta}_i$ is the vector containing the aggregated data values of δ_i. Equation (28) implies

$$\tilde{s}_{ij} \ = \ \gamma_i\gamma_j\tilde{s}_{yy} + \tilde{s}_{\delta_i\delta_j} + \gamma_i\tilde{s}_{y\delta_j} + \gamma_j\tilde{s}_{y\delta_i} \ , \tag{29}$$

where the empirical variances and covariances $\tilde{s}_{\delta_i\delta_j}$ and $\tilde{s}_{y\delta_i}$, $i, j = 1, \ldots, p$, are defined correspondingly to (3).

By part a) of the lemma, $\text{plim}_{n\to\infty}\tilde{s}_{yy} = \sigma_{yy}$. In Schmid *et al.* (2005a) it was also shown that $\text{plim}_{n\to\infty}\tilde{s}_{y\delta_i} = 0$. The probability limit of $\tilde{s}_{\delta_i\delta_j}$ is stated in the following corollary:

Corollary 1. *$\tilde{s}_{\delta_i\delta_j}$ converges in probability to $\sigma_{\delta_i\delta_j}/A$.*

Proof. Microaggregation by a sorting variable subdivides the set of indices $G := \{1,\ldots,n\}$ into groups $G_1,\ldots,G_k,\ldots,G_{n/A}$. Now, as $\bar{\bar{\delta}}_i = \bar{\delta}_i$ and $\text{plim}_{n\to\infty}\bar{\delta}_i = 0$ for $i = 1,\ldots,p$, $\text{plim}_{n\to\infty}\tilde{s}_{\delta_i\delta_j}$ can be written as

$$\begin{aligned}
\text{plim}_{n\to\infty}\tilde{s}_{\delta_i\delta_j} &= \text{plim}_{n\to\infty}\left(\frac{A}{n}\sum_{k=1}^{n/A}\Big(\frac{1}{A}\sum_{z\in G_k}\delta_{zi}\Big)\Big(\frac{1}{A}\sum_{z\in G_k}\delta_{zj}\Big)\right)\\
&= \text{E}\left(\sum_{z\in G_1}\delta_{zi}/A\sum_{z\in G_1}\delta_{zj}/A\right)\\
&= \frac{1}{A}\sigma_{\delta_i\delta_j}\ . \qquad (30)
\end{aligned}$$

□

From (29) and (30) we obtain

$$\tilde{\sigma}_{ij} := \text{plim}_{n\to\infty}\tilde{s}_{ij} = \gamma_i\gamma_j\sigma_{yy} + \frac{1}{A}\sigma_{\delta_i\delta_j}\ . \qquad (31)$$

As

$$\sigma_{ij} = \gamma_i\gamma_j\sigma_{yy} + \sigma_{\delta_i\delta_j} \qquad (32)$$

and $\gamma_i = \sigma_{iy}/\sigma_{yy}$, $i = 1,\ldots,p$, we finally obtain

$$\begin{aligned}
\tilde{\sigma}_{ij} &= \Big(1-\frac{1}{A}\Big)\gamma_i\gamma_j\sigma_{yy} + \frac{1}{A}\sigma_{ij}\\
&= \Big(1-\frac{1}{A}\Big)\frac{\sigma_{iy}\sigma_{jy}}{\sigma_{yy}} + \frac{1}{A}\sigma_{ij}\ , \qquad (33)
\end{aligned}$$

from which part c) follows.

Proof of Lemma 2:

For the proofs of parts b) and c), we refer to Schmid *et al.* (2005b). As to the proof of a), (29) and a corresponding equation for the non-aggregated data values yield

$$\begin{aligned}\sqrt{n}(\tilde{s}_{ij} - s_{ij}) &= \sqrt{n}\gamma_i\gamma_j(\tilde{s}_{yy} - s_{yy}) + \sqrt{n}\gamma_i(\tilde{s}_{y\delta_j} - s_{y\delta_j}) \\ &\quad + \sqrt{n}\gamma_j(\tilde{s}_{y\delta_i} - s_{y\delta_i}) + \sqrt{n}(\tilde{s}_{\delta_i\delta_j} - s_{\delta_i\delta_j}) \ . \end{aligned} \tag{34}$$

In Schmid *et al.* (2005b) it was shown that $\sqrt{n}(\tilde{s}_{yy} - s_{yy})$, $\sqrt{n}(\tilde{s}_{y\delta_j} - s_{y\delta_j})$, and $\sqrt{n}(\tilde{s}_{y\delta_i} - s_{y\delta_i})$ all converge in probability to 0. Therefore

$$\tilde{s}_{ij} - s_{ij} \ \sim \ \tilde{s}_{\delta_i\delta_j} - s_{\delta_i\delta_j} \tag{35}$$

and consequently

$$\tilde{s}_{ij} - \tilde{\sigma}_{ij} \ \sim \ s_{ij} + \frac{1}{A}s_{\delta_i\delta_j} - s_{\delta_i\delta_j} + \left(\tilde{s}_{\delta_i\delta_j} - \frac{1}{A}s_{\delta_i\delta_j}\right) - \tilde{\sigma}_{ij} \ . \tag{36}$$

Denote by $\hat{s}_{\delta_i\delta_j}$ the empirical variances and covariances of the *estimated* residuals $\hat{\delta}_{zi}$ and $\hat{\delta}_{zj}$, $1 \leq i, j, \leq p$, $z = 1, \ldots, n$, based on the non-aggregated data. Now, as $s_{\delta_i\delta_j} \sim \hat{s}_{\delta_i\delta_j}$ (compare Schmid *et al.* (2005b)), we have

$$\tilde{s}_{ij} - \tilde{\sigma}_{ij} \ \sim \ s_{ij} - \left(1 - \frac{1}{A}\right)\hat{s}_{\delta_i\delta_j} - \tilde{\sigma}_{ij} + \left(\tilde{s}_{\delta_i\delta_j} - \frac{1}{A}s_{\delta_i\delta_j}\right) \ . \tag{37}$$

Lemma 2a) now follows from (37), Lemma 1c), and from the fact that

$$\hat{s}_{\delta_i\delta_j} \ = \ s_{ij} - \frac{s_{iy}s_{jy}}{s_{yy}} \ . \tag{38}$$

Proof of Lemma 3: Using the same notation as in the proof of Corollary 1, we obtain

$$\begin{aligned}\sqrt{n}\,\Delta_{ij} &\approx \frac{A}{\sqrt{n}}\sum_{k=1}^{n/A}\Big(\frac{1}{A}\sum_{z\in G_k}\delta_{zi}\Big)\Big(\frac{1}{A}\sum_{v\in G_k}\delta_{vj}\Big) - \frac{1}{A\sqrt{n}}\sum_{k=1}^{n/A}\sum_{z\in G_k}\delta_{zi}\delta_{zj} \\ &= \frac{1}{A\sqrt{n}}\sum_{k=1}^{n/A}\sum_{\substack{z,v\in G_k\\ z\neq v}}\delta_{zi}\delta_{vj} \\ &= \sqrt{\frac{A}{n}}\sum_{k=1}^{n/A}\Delta_{ij(k)}\ , \end{aligned} \tag{39}$$

where "$\approx$" means that the difference converges to 0. The expressions $\Delta_{ij(k)} := \sum_{z,v\in G_k, z\neq v}\delta_{zi}\delta_{vj}/A^{3/2}$, $k = 1,\ldots,n/A$, are i.i.d. random variables with zero mean. By the central limit theorem, the $\sqrt{n}\Delta_{ij}$ are asymptotically jointly normally distributed. Moreover, the asymptotic covariance of $\sqrt{n}\Delta_{ij}$ and $\sqrt{n}\Delta_{mn}$ is equal to $\mathrm{E}(\Delta_{ij(1)}\Delta_{mn(1)})$. Now

$$\mathrm{E}(\Delta_{ij(1)}\Delta_{mn(1)}) = \frac{1}{A^3}\sum_{\substack{z,u,v,w\in G_1\\ z\neq u, v\neq w}}\mathrm{E}(\delta_{zi}\delta_{uj}\delta_{vm}\delta_{wn})\ . \tag{40}$$

Obviously, only the terms where $z = v$ and $u = w$ or where $z = w$ and $u = v$ contribute to the sum on the right hand side of (40). The number of these terms is $A(A-1)$ in both cases. Therefore

$$\sigma_{\Delta_{ij}\Delta_{mn}} = \frac{1}{n}\mathrm{E}(\Delta_{ij(1)}\Delta_{mn(1)}) = \frac{1}{n}\frac{(A-1)}{A^2}\left(\sigma_{\delta_i\delta_m}\sigma_{\delta_j\delta_n} + \sigma_{\delta_i\delta_n}\sigma_{\delta_j\delta_m}\right)\ . \tag{41}$$

Acknowledgements

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

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.

Laszlo, M. and S. Mukherjee (2005): "Minimum Spanning Tree Partitioning Algorithm for Microaggregation," IEEE Transactions on Knowledge and Data Engineering, 17, No. 7, 902-911.

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

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

Schmid, M. and H. Schneeweiss (2005): "The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study". In *Econometrics of Anonymized Micro Data*, ed. by W. Pohlmeier, G. Ronning, and J. Wagner. Jahrbücher für Nationalökonomie und Statistik, 225, No. 5, Stuttgart: Lucius & Lucius.

Schmid, M., H. Schneeweiss and H. Küchenhoff (2005a): "Consistent Estimation of a Simple Linear Model under Microaggregation," Discussion Paper 415, SFB 386, Department of Statistics, University of Munich.

Schmid, M., H. Schneeweiss and H. Küchenhoff (2005b): "Statistical Inference in a Simple Linear Model under Microaggregation," Discussion Paper 416, SFB 386, Department of Statistics, University of Munich.