Schmid, Matthias; Schneeweiss, Hans; Küchenhoff, Helmut

Working Paper

Consistent estimation of a simple linear model under microaggregation

Discussion Paper, No. 415

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

Suggested Citation: Schmid, Matthias; Schneeweiss, Hans; Küchenhoff, Helmut (2005) : Consistent estimation of a simple linear model under microaggregation, Discussion Paper, No. 415, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München, https://doi.org/10.5282/ubm/epub.1784

This Version is available at:
https://hdl.handle.net/10419/31051

Consistent Estimation of a Simple Linear Model Under Microaggregation

Matthias Schmid, Hans Schneeweiss and Helmut Küchenhoff

Department of Statistics, University of Munich

Ludwigstr. 33, 80539 München, Germany

Abstract

A problem statistical offices are increasingly faced with is guaranteeing confidentiality when releasing microdata sets. One method to provide safe microdata to is to reduce the information content of a data set by means of masking procedures. A widely discussed masking procedure is microaggregation, a technique where observations are grouped and replaced with their corresponding group means. However, while reducing the disclosure risk of a data file, microaggregation also affects the results of statistical analyses. The paper deals with the impact of microaggregation on a simple linear model. We show that parameter estimates are biased if the dependent variable is used to group the data. It turns out that the bias of the slope parameter estimate is a non-monotonic function of this parameter. By means of this non-monotonic relationship we develop a method for consistently estimating the model parameters.

Keywords: Microaggregation, simple linear model, bias, consistent estimation, disclosure control

1 Introduction

Over the last decades the development of empirical research in social and economic sciences has led to an increasing demand on microdata. However, with the growing

availability of databases, problems concerning data security have arisen: On the one hand, data protection laws demand that the data sets, most often containing sensitive information, have to be treated confidentially by the data collecting institutions. On the other hand, scientists need a maximum amount of information to draw the right conclusions from the data. Evidently, there is a trade-off between guaranteeing confidentiality and providing sufficient information to the researcher. This is what is commonly referred to as the *statistical disclosure control problem.*

One possibility to deal with this problem is the creation of factually anonymous microdata sets, also called scientific-use files. "Factually anonymous" means that the data user has to employ "an excessive amount of time, expenses, and manpower to allocate the data to the respondent" (Köhler (1999)). Clearly, factual anonymity implies that the information content of a data set has to be reduced to a certain extent. To achieve this, a rich variety of procedures has been developed, see Brand (2000) or Gottschalk (2004) for an overview. As each of these procedures may have an effect on data analysis, statistical research is confronted with the problem of investigating the impact of anonymization techniques on parameter estimation, hypothesis testing, etc.

In this paper, we focus on microaggregation, a widely discussed anonymization procedure for continuous data (Anwar (1993), Defays and Nanopoulos (1993), Defays and Anwar (1998), Domingo-Ferrer and Mateo-Sanz (2002), Lechner and Pohlmeier (2003), Rosemann (2004)). The main idea of microaggregation is to group the observations in a data set and replace the original data values with their corresponding group means. The various types of microaggregation procedures mainly differ in how

the grouping of the data is done. Usually, a similarity criterion such as the Euclidean distance or the Mahalanobis distance is used to form the groups.

The microaggregation technique considered in this paper uses a so-called "leading variable" to group the data (Paass and Wauschkuhn (1985), Mateo-Sanz and Domingo-Ferrer (1998)). The leading variable can either be one of the regressors or the dependent variable in a statistical model. Groups are then formed by data records having similar values for the leading variable. Throughout this paper, a fixed group size (also called "aggregation level") is used.

We want to study the effect of this type of microaggregation on the estimation of a simple linear regression in continuous variables. It is well-known that microaggregation with respect to one (or several) exogenous variables as well as random microaggregation have no effect on the unbiasedness property of OLS, see Feige and Watts (1972) or Lechner and Pohlmeier (2003). What seems to be less well-known is that microaggregation with respect to the endogenous variable does have an effect (but see Feige and Watts (1972), who hint at the possibility of such an effect, however, without investigating it in any detail). The purpose of this paper is to study this effect, in particular the magnitude of the bias resulting from this kind of microaggregation. By analyzing the relation between bias and model parameters, we can then construct a consistent estimator of the slope parameter.

It turns out that the aggregation bias of the OLS of the slope parameter β in a simple linear regression model depends on the error variance of the model and on the slope parameter itself. Contrary to the well-known attenuation effect of measurement error models, the OLS bias of the slope parameter is always positive for an ascending line

and negative for a descending line. It is zero when the line is flat and again tends to zero when the slope becomes infinite. The bias is thus a non-monotonic function of β. The relative bias of OLS is, for $\beta > 0$, a monotonically decreasing function of the correlation between the dependent variable and the regressor. These results are proved and made plausible in the following sections. Furthermore, the behavior of the OLS estimator for finite samples is examined by means of a systematic simulation study.

In section 2, we start with a summary of the results concerning microaggregation with respect to the exogenous variable X. In section 3, we illustrate the effect of microaggregation with respect to the endogenous variable Y on a linear model. This is done by discussing a very simple situation that involves a discrete error structure. Section 4 contains theoretical results on the effects of microaggregation with respect to Y on a linear model with normally distributed errors. Furthermore, a method for correcting the aggregation bias is developed. In section 5, a systematic simulation study is carried out. Section 6 contains a concluding summary. Proofs are relegated to the appendix. Further results concerning t-tests and the effect of microaggregation on the variance of the OLS estimator of β will be presented in a subsequent paper.

2 Microaggregation with Respect to X

As stated in the introduction we want to investigate the impact of microaggregation on the parameter estimates of the simple linear model

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

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

Suppose we have an i.i.d. sample of size n and two vectors $y := (y_1, ...y_n)'$, $x := (x_1, ..., x_n)'$ containing the data values. Denote by $e := (\epsilon_1, \ldots, \epsilon_n)'$ the error vector having independent and identically normally distributed components.

Now, two possibilities of microaggregating the data exist:

A) The data can be aggregated with respect to the covariate X. As mentioned before, this type of microaggregation (using X as the leading variable) has already been investigated by Feige and Watts (1972) and Lechner and Pohlmeier (2003). In this section, we briefly discuss their results. In addition, we show that β and σ_ϵ^2 can be consistently estimated by the naive least squares estimates.

B) The data can be aggregated with respect to the dependent variable Y. This procedure (where Y is the leading variable) has not been studied in the literature yet. In sections 3 - 5 we investigate the impact of this type of microaggregation.

Let us now explain how a data set is aggregated with respect to the covariate X: First of all, the data set has to be ordered according to the magnitude of X. We say for short that the data are "sorted by X". After having chosen an aggregation level A, the sorted data set is subdivided into n/A groups, each consisting of A adjacent data values. For simplicity, we assume that n is a multiple of A. In each group, the

data (x_i, y_i), $i = 1, \ldots, n$, are averaged and the averages are assigned to the items of the group. The application of this procedure to the data is the same as multiplying the sorted vectors $y_{sort(x)}$, $x_{sort(x)}$ with an idempotent matrix D consisting of ones and zeroes:

$$D := \frac{1}{A} \cdot \begin{pmatrix} 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ & \vdots & & \ddots & & \vdots & \\ 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \\ \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \underbrace{0 \quad \cdots \quad 0}_{A} & & & \cdots & 1 & \cdots & 1 \end{pmatrix}. \tag{2}$$

Denote by $\tilde{y}_x$ and $\tilde{x}_x$ the vectors containing the data that have been aggregated with respect to the random variable X. These vectors can now be written as $\tilde{y}_x = Dy_{sort(x)}$ and $\tilde{x}_x = Dx_{sort(x)}$. Similarly, the aggregated design matrix $\tilde{X}_x := (\mathbf{1}, \tilde{x}_x)$ can be written as $\tilde{X}_x = D \cdot (\mathbf{1}, x_{sort(x)})$. It is easily seen that by aggregating the data with respect to X, the means of the original variables in the data set are preserved. Moreover, the aggregation procedure does not depend on the error structure of the linear model, implying that the least squares estimate

$$\tilde{\gamma} = (\tilde{\alpha}, \tilde{\beta})' := (\tilde{X}_x' \tilde{X}_x)^{-1} \tilde{X}_x' \tilde{y}_x \tag{3}$$

becomes unbiased, see Feige and Watts (1972) or Lechner and Pohlmeier (2003) (we implicitely assume that $\tilde{X}_x' \tilde{X}_x$ is nonsingular). Moreover, the following theorem holds:

Theorem 1. *The estimate $\tilde{\gamma}$ based on the microaggregated data is a consistent estimate of γ.*

Proof: See appendix A.

Theorem 1 shows that $\tilde{\gamma}$ remains a consistent estimator of γ, just as the least squares estimator $\hat{\gamma}$ computed from the original data. But there is a loss of efficiency as $\tilde{\gamma}$ has greater variance than $\hat{\gamma}$, see Feige and Watts (1972) or Lechner and Pohlmeier (2003). This loss of efficiency, however, tends to zero with increasing n, so that asymptotically $\tilde{\gamma}$ and $\hat{\gamma}$ are equally efficient.

The next result concerns the residual sum of squares $\tilde{e}_x'\tilde{e}_x := (\tilde{y}_x - \tilde{X}_x\tilde{\gamma})'(\tilde{y}_x - \tilde{X}_x\tilde{\gamma})$:

Theorem 2. (A/n) $\tilde{e}_x'\tilde{e}_x$ *is a consistent estimator of* σ_ϵ^2.

Proof: See appendix A.

Theorem 2 shows that the "naive" variance estimate $\tilde{\sigma}_\epsilon^2 := 1/n\ (\tilde{y}_x - \tilde{X}_x\tilde{\gamma})'(\tilde{y}_x - \tilde{X}_x\tilde{\gamma})$ does not converge to the true residual variance σ_ϵ^2. However, by multiplying $\tilde{\sigma}_\epsilon^2$ with A, one can easily obtain a consistent estimate of σ_ϵ^2.

Note that Theorems 1 and 2 also hold if a linear model with p predictors is considered and a leading variable X_k, $1 \leq k \leq p$, is used for microaggregation.

3 Microaggregation with Respect to Y - Analysis of a Linear Model with Discrete Errors

In this and the remaining sections, we exclusively study what happens to model (1) if the data have been aggregated with respect to Y. Aggregation with respect to Y is

carried out in the same way as described in section 2 for the analogous aggregation procedure with respect to X: After the data set has been sorted by the leading variable Y, the vectors $\tilde{y}_y$ and $\tilde{x}_y$ containing the aggregated data values become $\tilde{y}_y = Dy_{sort(y)}$ and $\tilde{x}_y = Dx_{sort(y)}$.

In the following and contrary to the notation of section 2, $\tilde{\gamma}$ denotes the least squares estimate of γ computed from the data that have been microaggregated with respect to Y (and not with respect to X as in section 2). Again, $\tilde{\gamma}$ can be written as

$$\begin{aligned} \tilde{\gamma} &= (\tilde{X}_y'\tilde{X}_y)^{-1}\tilde{X}_y'\tilde{y}_y \\ &= (X_{sort(y)}'DX_{sort(y)})^{-1}X_{sort(y)}'Dy_{sort(y)} \ , \end{aligned} \tag{4}$$

where $\tilde{X}_y := (\mathbf{1}, \tilde{x}_y)$. $X_{sort(y)}$ denotes the design matrix after sorting the data with respect to Y.

The main difference to situation A in section 2 is that $X_{sort(y)}$ now depends on the error structure of the respective linear model. This means that the results described in section 2 can not be applied to $\tilde{\gamma}$. It is not obvious at all how to compute $\mathrm{E}(\tilde{\gamma})$ and how to assess whether $\tilde{\gamma}$ is unbiased or not.

In order to get a first idea of the effect of microaggregation with respect to Y, we start by studying a very simple (artificial) linear model involving a discrete error structure and a discrete regressor X. Let the vector x be given by $(1, \ldots, 8, 1, \ldots, 8)'$ and consider the deterministic vector of residuals $e = (0.5, \ldots, 0.5, -0.5, \ldots, -0.5)'$. Assuming α to be zero, the response vector y becomes

$$y = (\beta \cdot 1 + 0.5, \ldots, \beta \cdot 8 + 0.5, \ldots, \beta \cdot 1 - 0.5, \ldots, \beta \cdot 8 - 0.5)' \ . \tag{5}$$

Fig. 1 shows the resulting plot of y vs. x for $\beta = 1$.

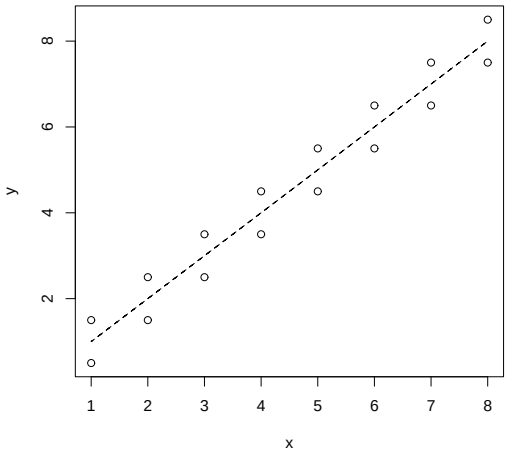

Figure 1: Plot of y vs. x ($\beta = 1$)

Let us now see what happens if the data are aggregated with respect to Y (in the following, we use an aggregation level of $A = 2$). Figs. 2 - 4 show the effects of the aggregation step by step for five values of β (β= 0.05, β=0.18, β=0.25, β=0.5, and β=1.5). The filled-in dots represent the aggregated data values.

As long as β is close to zero, aggregating the data set with respect to Y is the same as aggregating the points lying below the true regression line and the points lying above the true regression line separately. As the order of $(x_1, \ldots, x_8)'$ and $(x_9, \ldots, x_{16})'$ is the same as the order of $(y_1, \ldots, y_8)'$ and $(y_9, \ldots, y_{16})'$ respectively, the least squares estimate $\tilde{\beta}$ is unbiased, and the estimated regression line based on the aggregated data values is the same as the true regression line (Fig. 2).

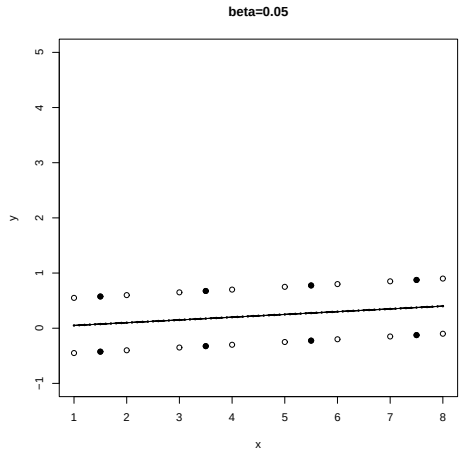

Figure 2: Plot of y vs. x and $\tilde{y}_y$ vs. $\tilde{x}_y$ for $\beta = 0.05$ (dotted line = true regression)

Fig. 3 ($\beta = 0.18$) shows a different picture: As β increases, the "middle" points $(1, \beta \cdot 1 + 0.5)$, $(7, \beta \cdot 7 - 0.5)$ and $(2, \beta \cdot 2 + 0.5)$, $(8, \beta \cdot 8 - 0.5)$ are grouped, forcing the corresponding aggregated data values to move in the direction of $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$. It is well known from linear model theory that extreme data values situated far from $\bar{x}$ (in this case, $(1.5, \frac{1}{2}(\beta \cdot 1 - 0.5 + \beta \cdot 2 - 0.5))$ and $(7.5, \frac{1}{2}(\beta \cdot 7 + 0.5 + \beta \cdot 8 + 0.5))$) have a big influence on the slope of the estimated regression line. This is why $\tilde{\beta}$ in Fig. 3 has a positive bias.

The above described effect becomes even stronger if β continues to increase (Fig. 3, $\beta = 0.25$): Again, two aggregated data values move in the direction of $\bar{x}$, causing the bias of $\tilde{\beta}$ to increase even more.

However, as the true regression line becomes steeper, the number of points lying exactly on the true regression line increases, too (two points if $\beta = 0.18$, four points if $\beta = 0.25$). This has an adverse effect on the estimate of the slope: The bias of $\tilde{\beta}$

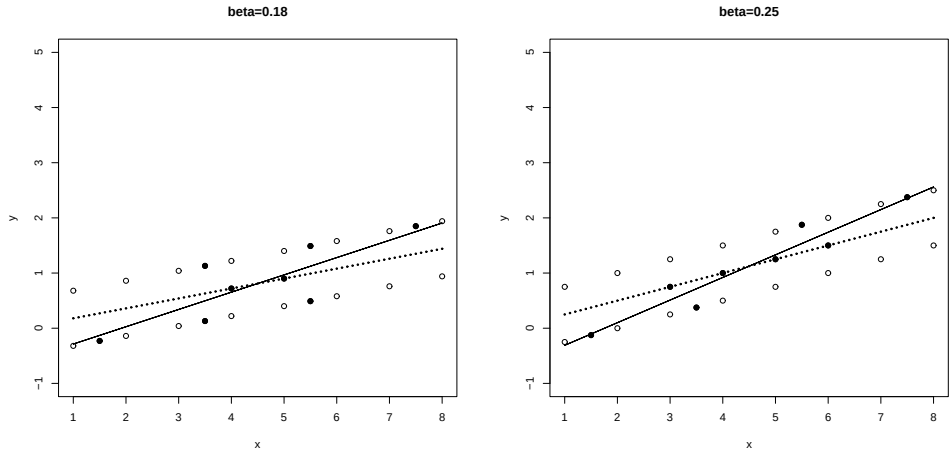

Figure 3: Plot of y vs. x and $\tilde{y}_y$ vs. $\tilde{x}_y$ for $\beta = 0.18$ and $\beta = 0.25$ (dotted line = true regression)

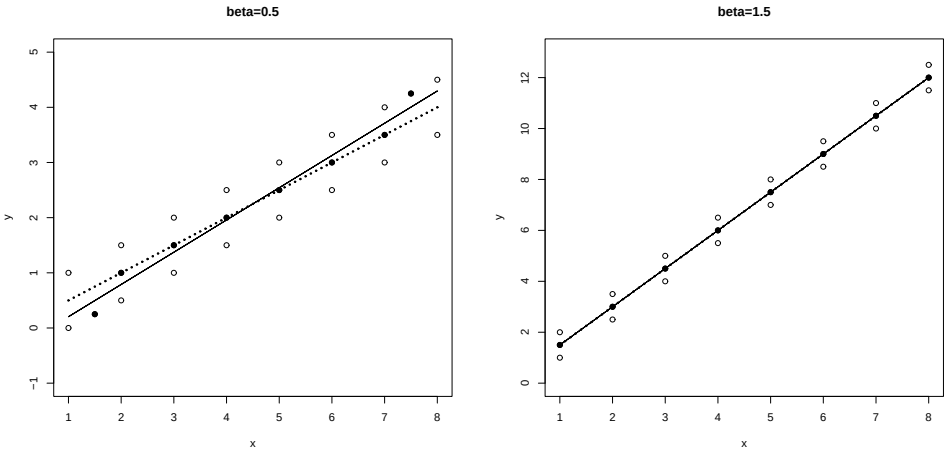

Figure 4: Plot of y vs. x and $\tilde{y}_y$ vs. $\tilde{x}_y$ for $\beta = 0.5$ and $\beta = 1.5$ (dotted line = true regression)

begins to decline as more and more aggregated data values lie on the true regression line (Fig. 4, $\beta = 0.5$).

Finally, as β goes to infinity, *all* aggregated data values lie on the true regression

line, and $\tilde{\beta}$ equals the true β again. This can be seen from Fig. 4 ($\beta = 1.5$) where the two regression lines have become identical, just like in Fig. 2.

Thus we can conclude that the bias of $\tilde{\beta}$ is zero as long as β is close to zero. As the values of β increase, bias($\tilde{\beta}$) becomes positive at first. As $\beta \to \infty$, bias($\tilde{\beta}$) declines and becomes zero again. Fig. 5 illustrates this result. It is also clear from Figs. 2 - 5 that for negative values of β, bias($\tilde{\beta}$) becomes negative at first. As $\beta \to -\infty$, bias($\tilde{\beta}$) becomes zero again.

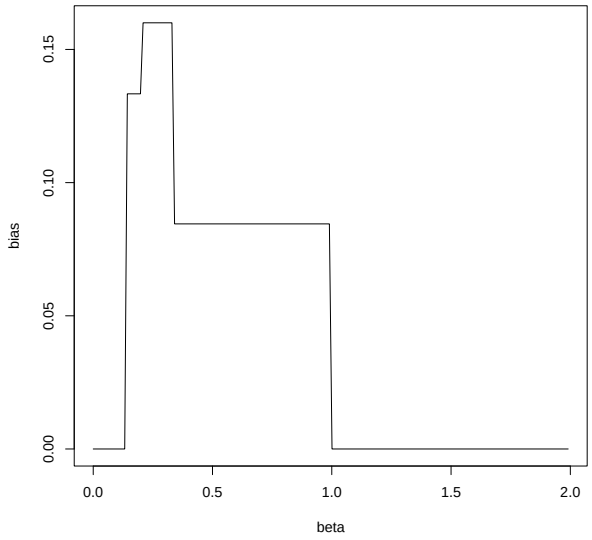

Figure 5: Plot of bias($\tilde{\beta}$) vs. β, $x = (1, \ldots, 8, 1, \ldots, 8)'$

Fig. 6 shows what happens if the sample size n is increased (here, the first half of x is $(1, 1.02, 1.04, 1.06, \ldots, 24)'$): The bias of the least squares estimate $\tilde{\beta}$ is almost a smooth function of β.

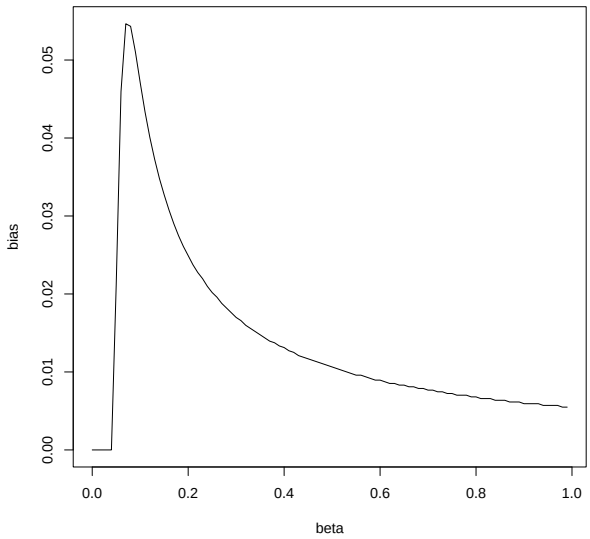

Figure 6: Plot of bias($\tilde{\beta}$) vs. β, first half of x equals $(1, 1.02, 1.04, 1.06, \ldots, 24)'$

Finally, we modify the above model by replacing the deterministic residuals with a simple stochastic error structure: Let the vector x be $(1, 2, 3, 5)'$. The residuals $\epsilon_1, \ldots, \epsilon_4$ are now assumed to take on the values $+0.5$ or -0.5, each with probability $1/2$. As there are $2^4 = 16$ possible values for the vector $e = (\epsilon_1, \ldots, \epsilon_4)'$, the mean of $\tilde{\beta}$ can be computed by averaging the 16 least squares estimates for each value of β. The resulting bias curve shown in Fig. 7 is very similar to Fig. 5, and the conclusions concerning the deterministic-error model can be applied to the above stochastic-error model as well. In the next section, we show that the results derived in this section also hold for a linear model with normally distributed variables.

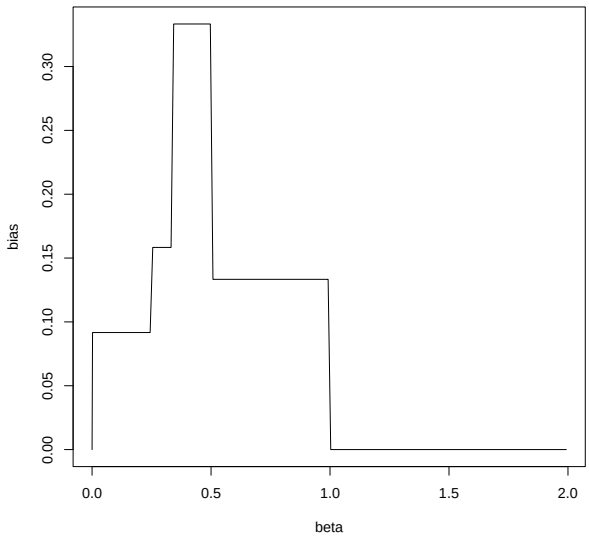

Figure 7: Plot of bias($\tilde{\beta}$) vs. β, $x = (1, 2, 3, 5)'$

4 Microaggregation with Respect to Y - Analysis of a Linear Model with Normally Distributed Variables

4.1 Bias of the Slope Parameter

In this section, we again consider model (1). X is now assumed to follow a normal distribution with mean μ_x and variance σ_x^2. The error variable ϵ is assumed to be normally distributed with zero mean and variance σ_ϵ^2. Assuming X and ϵ to be independent, it follows that Y is normally distributed as well with mean $\mu_y := \alpha + \beta\mu_x$ and variance $\sigma_y^2 := \beta^2\sigma_x^2 + \sigma_\epsilon^2$. Denote by ρ the correlation coefficient between X and Y.

Suppose we have n independent and identically distributed observations $(x, y) := (x_i, y_i)_{i=1,\dots,n}$. Again, we study what happens to the least squares estimate $\tilde{\beta}$ if the data are microaggregated with respect to the response variable Y. Our main concern is in the asymptotic properties of $\tilde{\beta}$.

First note that $\tilde{\beta}$ is given by

$$\tilde{\beta} = \frac{S_{\tilde{x}_y \tilde{y}_y}}{S^2_{\tilde{x}_y}} \ , \tag{6}$$

where

$$S_{\tilde{x}_y \tilde{y}_y} := \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{x}_{y,i} - \bar{\tilde{x}}_y\right)\left(\tilde{y}_{y,i} - \bar{\tilde{y}}_y\right) \tag{7}$$

is the empirical covariance of $\tilde{x}_y$ and $\tilde{y}_y$ and

$$S^2_{\tilde{x}_y} := \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{x}_{y,i} - \bar{\tilde{x}}_y\right)^2 \tag{8}$$

is the empirical variance of $\tilde{x}_y$. By investigating the asymptotic behavior of $S_{\tilde{x}_y \tilde{y}_y}$ and $S^2_{\tilde{x}_y}$, we can analyze the asymptotic behavior of $\tilde{\beta}$.

Lemma 1. *Denote by $S^2_{\tilde{y}_y}$ the empirical variance of $\tilde{y}_y$. Then the following results hold:*

a) $S^2_{\tilde{y}_y}$ *converges to* σ_y^2 *in probability.*

b) $S^2_{\tilde{x}_y}$ *converges in probability to* $\tilde{\sigma}_x^2 := \sigma_x^2 / f(\rho)$*, where*

$$f(\rho) := \frac{1}{\frac{1}{A} + \left(1 - \frac{1}{A}\right)\rho^2} \ . \tag{9}$$

c) $S_{\tilde{x}_y \tilde{y}_y}$ *converges in probability to* $\sigma_{xy} := \rho \sigma_x \sigma_y$.

Proof: See appendix B.

Now the following theorem holds:

Theorem 3. *$\tilde{\beta}$ converges in probability to $\beta f(\rho)$.*

Proof: From Lemma 1 we have

$$\tilde{\beta} = \frac{S_{\tilde{x}_y \tilde{y}_y}}{S^2_{\tilde{x}_y}} \quad \rightarrow \quad \frac{\sigma_{xy}}{\sigma_x^2 / f(\rho)} = \beta f(\rho) \ . \tag{10}$$

We see that for $\beta \neq 0$, the asymptotic relative bias of $\tilde{\beta}$ is equal to $f(\rho)$. It follows from (9) that $f(\rho) > 1$ for $\rho \neq 1$ and $A > 1$. Thus, $|\beta|$ is systematically *overestimated* by $\tilde{\beta}$, at least for large n. If $\beta = 0$, $\tilde{\beta}$ becomes a consistent estimate of β despite the microaggregation of the data.

Fig. 8 shows the graph of $f(\rho)$. The aggregation level A was set to three. We see that if $\rho = 0$, $f(\rho) = A$. Furthermore, $f(\rho) \rightarrow 1$ if $|\rho| \rightarrow 1$. This means that for large values of $|\rho|$, the bias of $\tilde{\beta}$ disappears. This is a very plausible result because a large value of $|\rho|$ implies that sorting the data with respect to Y is approximately the same as sorting the data with respect to X. As $\tilde{\beta}$ is unbiased in case of aggregating the data with respect to X, the least squares estimate based on the data that have been aggregated with respect to Y should be (at least approximately) unbiased, too, if $|\rho|$ is large.

Noting that

$$\rho^2 = \rho^2(\beta, \sigma_x^2, \sigma_\epsilon^2) = \frac{\beta^2}{\beta^2 + \sigma_\epsilon^2 / \sigma_x^2} \ , \tag{11}$$

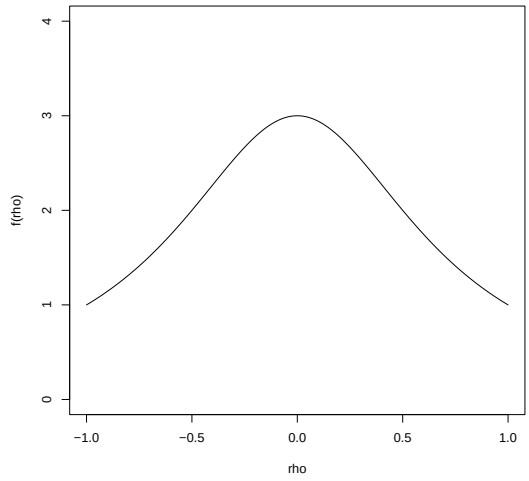

Figure 8: Plot of $f(\rho)$ vs. ρ, A=3

we can express the asymptotic relative bias of $\tilde{\beta}$ as a function of β and σ_ϵ^2:

$$f(\rho) = \frac{A(\beta^2 + v^2)}{A\beta^2 + v^2} \ , \tag{12}$$

where $v^2 := \sigma_\epsilon^2/\sigma_x^2$. Similarly, the asymptotic bias of $\tilde{\beta}$, $b := \operatorname{plim} \tilde{\beta} - \beta$, is found to be

$$\begin{aligned} b &= \beta(f(\rho) - 1) \\ &= (A-1) \cdot \frac{\beta}{1 + \frac{A}{v^2}\beta^2} \ . \end{aligned} \tag{13}$$

Thus, for small values of β, the bias grows approximately proportionally with β, whereas for large values of β it flattens to zero. It has its extreme values at $\beta_m = \pm v/\sqrt{A}$ with largest absolute bias $\frac{A-1}{2}\frac{v}{\sqrt{A}}$.

4.2 Bias of the Intercept

Concerning the estimation of the intercept α, the asymptotic bias a of the naive estimate $\tilde{\alpha} := \bar{\tilde{y}}_y - \tilde{\beta}\bar{\tilde{x}}_y$ can be evaluated as follows:

$$\begin{aligned} a &:= \operatorname{plim}\bigl(\bar{\tilde{y}}_y - \tilde{\beta}\bar{\tilde{x}}_y\bigr) - \alpha \\ &= \bigl(\mu_y - \beta f(\rho)\mu_x\bigr) - \bigl(\mu_y - \beta\mu_x\bigr) \\ &= -b\mu_x \ . \end{aligned} \tag{14}$$

Thus, if $\beta > 0$ and $\mu_x > 0$, $\tilde{\alpha}$ is asymptotically smaller than the true value of α.

4.3 Bias of the Residual Variance

Finally, we show what happens to the naive estimate $\tilde{\sigma}_\epsilon^2 = S_{\tilde{y}_y}^2 - \tilde{\beta}^2 S_{\tilde{x}_y}^2$ if $n \to \infty$.

By Lemma 1 and Theorem 3,

$$\begin{aligned} \operatorname{plim}\tilde{\sigma}_\epsilon^2 &= \sigma_y^2 - \beta^2 f(\rho)^2 \frac{\sigma_x^2}{f(\rho)} \\ &= \beta^2\sigma_x^2 + \sigma_\epsilon^2 - f(\rho)\beta^2\sigma_x^2 \\ &= \bigl(1 - f(\rho)\bigr)\beta^2\sigma_x^2 + \sigma_\epsilon^2 \\ &= \frac{v^2 + \beta^2}{v^2 + A\beta^2}\,\sigma_\epsilon^2 \\ &= \frac{1}{A}\,f(\rho)\sigma_\epsilon^2 \ . \end{aligned} \tag{15}$$

4.4 Bias Correction

We can use the bias formulas of the previous sections to correct for the bias of $\tilde{\beta}$, $\tilde{\alpha}$, and $\tilde{\sigma}_\epsilon^2$. Denote by $\tilde{\rho}$ the empirical correlation coefficient based on the aggregated data. Now,

$$\tilde{\rho}^2 = \frac{S^2_{\tilde{x}_y \tilde{y}_y}}{S^2_{\tilde{x}_y} S^2_{\tilde{y}_y}} \quad \stackrel{\text{plim}}{\longrightarrow} \quad \frac{\sigma^2_{xy}}{\sigma^2_x \sigma^2_y} f(\rho) = \rho^2 f(\rho) \ . \tag{16}$$

Therefore, we can find a consistent estimate $\tilde{\rho}_c^2$ by equating $\tilde{\rho}^2$ to $\rho^2 f(\rho)$:

$$\tilde{\rho}^2 = \tilde{\rho}_c^2 f(\tilde{\rho}_c) = \frac{A\tilde{\rho}_c^2}{1 + (A-1)\tilde{\rho}_c^2} \ . \tag{17}$$

Solving for $\tilde{\rho}_c^2$ yields

$$\tilde{\rho}_c^2 = \frac{\tilde{\rho}^2}{A - \tilde{\rho}^2(A-1)} \ . \tag{18}$$

From (10) and (18), we get a consistent estimate of β:

$$\tilde{\beta}_c = \frac{\tilde{\beta}}{f(\tilde{\rho}_c)} = \frac{\tilde{\beta}\big(1 + (A-1)\tilde{\rho}_c^2\big)}{A} = \frac{\tilde{\beta}}{A - (A-1)\tilde{\rho}^2} \ . \tag{19}$$

A consistent estimate of the intercept α can be obtained from (14) and (19):

$$\tilde{\alpha}_c = \tilde{\alpha} + (\tilde{\beta} - \tilde{\beta}_c)\bar{\tilde{x}}_y \ . \tag{20}$$

To derive a consistent estimate of σ_ϵ^2, we make use of (15). σ_ϵ^2 can be consistently estimated by

$$\tilde{\sigma}^2_{\epsilon,c} := \frac{A(S^2_{\tilde{y}_y} - \tilde{\beta}^2 S^2_{\tilde{x}_y})}{f(\tilde{\rho}_c)} \ . \tag{21}$$

5 Simulations

5.1 Bias of $\tilde{\beta}$ and $\tilde{\sigma}^2_\epsilon$ for Finite Samples

In this section, we check to which extent the asymptotic results of section 4 hold in realistic data situations. For this purpose, we carried out a systematic simulation study, setting the sample size n to 300. For each parameter combination (β, σ_ϵ), the bias of the least squares estimate $\tilde{\beta}$ was estimated from 500 randomly generated data sets (x_i, ϵ_i), $i = 1, \ldots, 300$. The random numbers x_i were drawn iid from a normal distribution with zero mean and variance $\sigma_x^2 = 4$. Note that it is sufficient to consider variations in β and σ_ϵ^2 only. σ_x^2 can be kept fixed without loss of generality. The aggregation level was chosen to be $A = 3$, α was set to one.

In Fig. 9, bias$(\tilde{\beta})$ is plotted vs. β for different values of σ_ϵ, together with the asymptotic bias b from (13). Obviously, the approximation of the finite sample bias by its asymptotic counterpart works very well. We see that the bias of $\tilde{\beta}$ becomes zero if β is zero. Moreover, bias$(\tilde{\beta}) \to 0$ as $|\beta|$ goes to infinity. If σ_ϵ gets larger, bias$(\tilde{\beta})$ gets larger as well. Note further the remarkable resemblance of Figs. 6 and 9: Obviously, the results derived for the simple model in section 3 can be applied to the much more realistic case of a linear model with normally distributed variables.

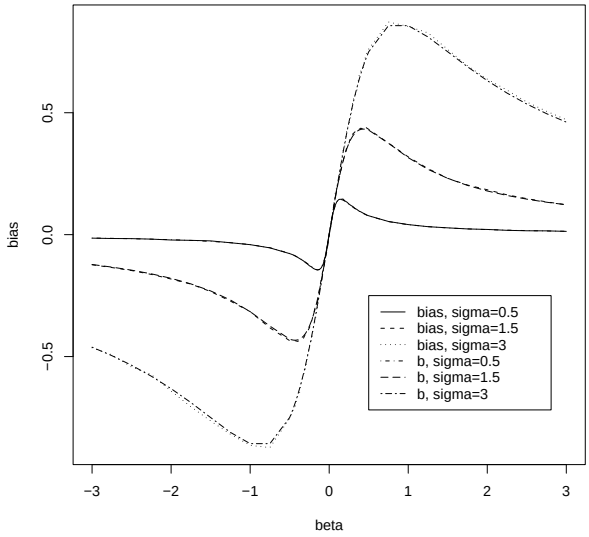

Figure 9: Bias of $\tilde{\beta}$ as a function of β for various values of σ_ϵ

In Fig. 10, the relationship between β, σ_ϵ, and bias($\tilde{\beta}$) is illustrated by means of a three dimensional plot. Again, we see that bias($\tilde{\beta}$) gets larger as σ_ϵ increases. Fig. 11 shows the relative bias of $\tilde{\beta}$ for various values of ρ, together with the asymptotic relative bias $f(\rho)$. We see that the approximation of the relative bias by $f(\rho)$ is very good.

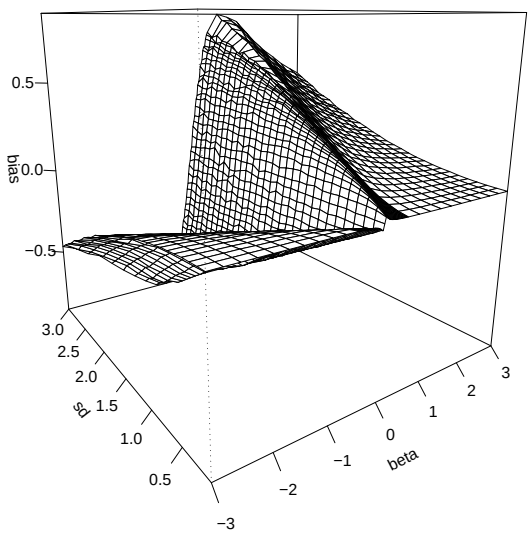

Figure 10: Bias of $\tilde{\beta}$ as a function of β and σ_ϵ (sd=σ_ϵ)

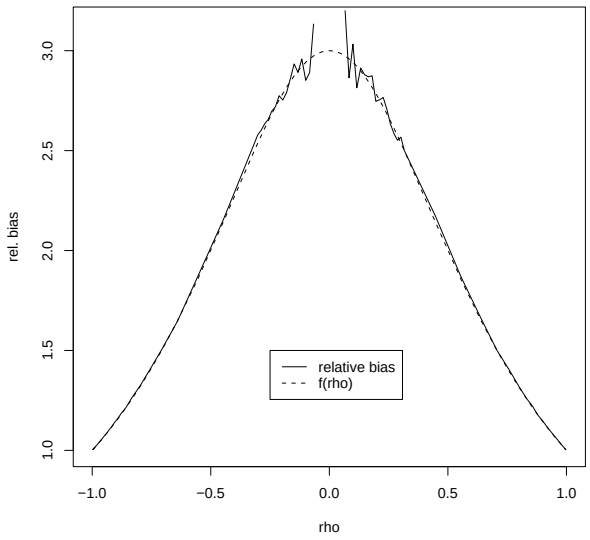

Figure 11: Relative bias of $\tilde{\beta}$ as a function of ρ

Fig. 12 shows what happens if the aggregation level A is altered (in the following, we set $\sigma_\epsilon = 3$). We see that $|\text{bias}(\tilde{\beta})|$ gets larger as A increases.

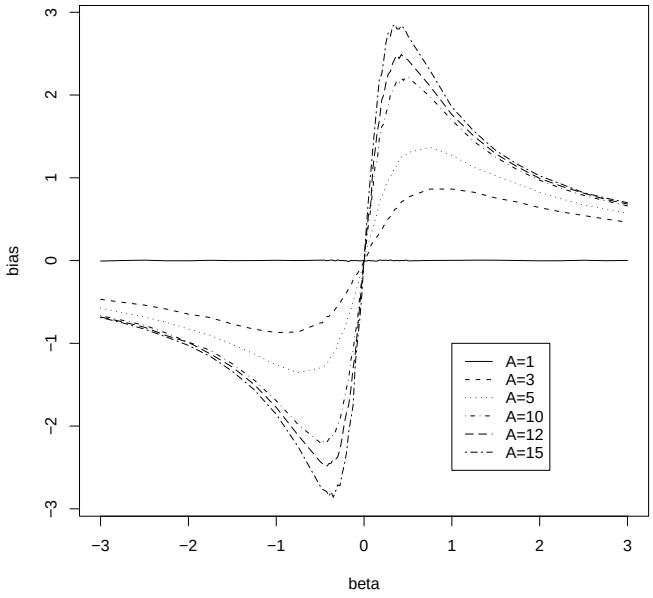

Figure 12: Bias of $\tilde{\beta}$ for various aggregation levels ($\sigma_\epsilon = 3$)

For n sufficiently large, the bias of $\tilde{\beta}$ should approach the asymptotic bias b and should therefore be essentially independent of n. Fig. 13 supports this result: n does not seem to have any influence on $\text{bias}(\tilde{\beta})$. The curves corresponding to $n = 50$, $n = 100$, $n = 150$, $n = 200$, $n = 250$, and $n = 300$ are almost identical. We see that even for small sample sizes, the approximation of $\text{bias}(\tilde{\beta})$ by the asymptotic bias b works well.

Fig. 14 shows the mean of the estimated residual standard deviations based on the aggregated data from the above simulation study. We also see the graph of

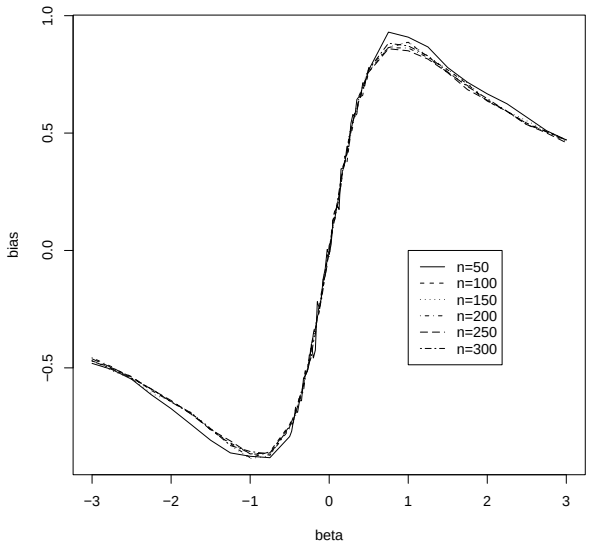

Figure 13: Bias of $\tilde{\beta}$ ($A = 3$, $\sigma_\epsilon = 3$)

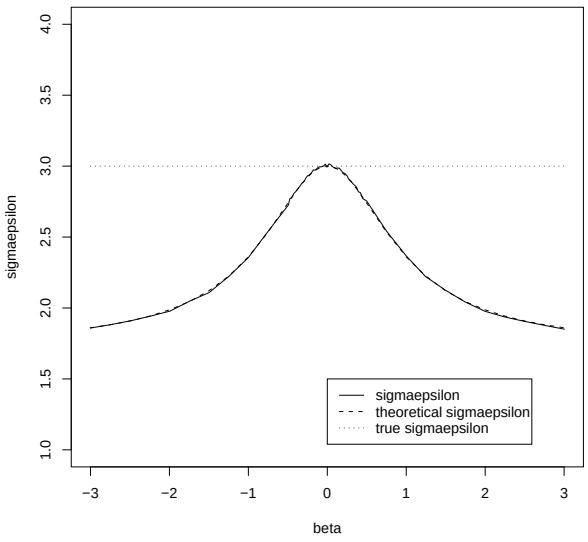

Figure 14: Residual standard deviation $\hat{\sigma}_\epsilon$, true $\sigma_\epsilon = 3$ ($A = 3$)

the function derived in (15). Again, the approximation of $\mathrm{E}(\tilde{\sigma}_\epsilon)$ by its asymptotic counterpart is very good.

By (15), if $\beta = 0$, $\operatorname{plim} \tilde{\sigma}_\epsilon^2 = \sigma_\epsilon^2$. This is plausible because in this case, $\sigma_\epsilon^2 = \sigma_y^2$ and $\operatorname{plim} S_{\tilde{y}_y}^2 = \sigma_y^2$ as well as $\operatorname{plim} \tilde{\beta} = \beta$.

We also see that as $\beta \to \pm\infty$, $\operatorname{plim} \tilde{\sigma}_\epsilon^2$ goes to $1/A\ \sigma_\epsilon^2$. Again, this is a plausible result: As $\beta \to \pm\infty$, aggregating with respect to Y is approximately the same as aggregating with respect to X. Therefore the residual variance estimate approximately takes the same value as when aggregation is performed with respect to X (see Theorem 2).

5.2 Finite Sample Bias of $\tilde{\beta}_c$ and $\tilde{\sigma}_{\epsilon,c}^2$

After having shown that the approximation of $\mathrm{bias}(\tilde{\beta})$ by the asymptotic bias b works very well in practice, we now investigate the behavior of the corrected estimator $\tilde{\beta}_c$ in realistic data situations. To achieve this, we computed the bias of $\tilde{\beta}_c$ for various n and various values of β. As before, we set $\sigma_\epsilon = 3$ and $A = 3$. Fig. 15 shows the bias of $\tilde{\beta}_c$, together with the bias of the estimator $\hat{\beta}$ based on the non-aggregated data. Obviously, if n is small, the bias of $\tilde{\beta}_c$ differs from its asymptotic bias (which is equal to zero). As n increases, the correction of $\tilde{\beta}$ works as it should: The mean of $\tilde{\beta}_c$ becomes almost identical to the true slope parameter.

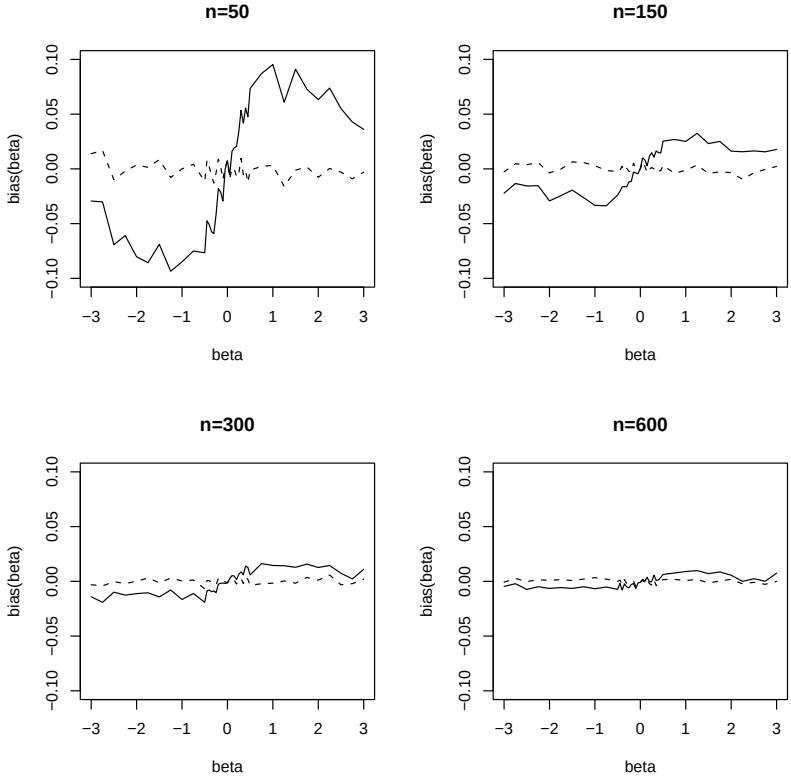

Figure 15: Bias curves of $\tilde{\beta}_c$ and $\hat{\beta}$ ($\sigma_\epsilon = 3$)

In the same way, we computed the bias of $\tilde{\sigma}^2_{\epsilon,c}$ for various n and various values of σ^2_ϵ. Here we set $\beta = 1$ and $A = 3$. Fig. 16 shows the bias of $\tilde{\sigma}^2_{\epsilon,c}$, together with the bias of the estimator $\hat{\sigma}^2_\epsilon$ based on the non-aggregated data set. We see that if n is small, the bias of the corrected estimator $\tilde{\sigma}^2_{\epsilon,c}$ severely differs from its asymptotic bias (which is equal to zero). As n increases, the correction of $\tilde{\sigma}^2_\epsilon$ works as it should: The mean of $\tilde{\sigma}^2_{\epsilon,c}$ is almost identical to the true residual variance.

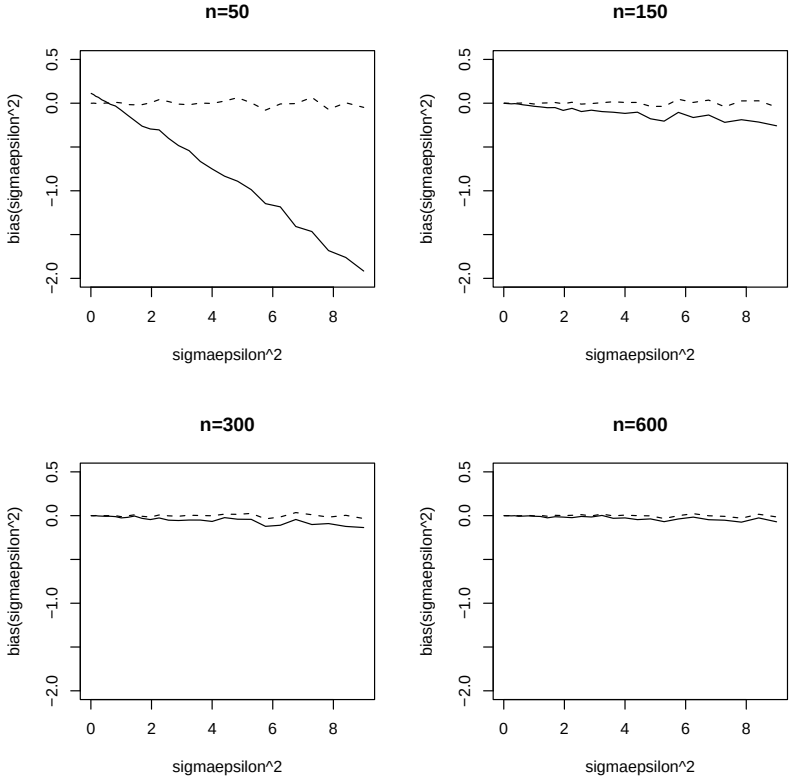

Figure 16: Bias curves of $\tilde{\sigma}^2_{\epsilon,c}$ and $\hat{\sigma}^2_{\epsilon}$ $(\beta = 1)$

6 Conclusion

As anonymization techniques and the creation of scientific-use files have become more and more important over the last ten years, it is necessary to study both disclosure risks and the impact of anonymization techniques on statistical analysis. In this paper, we focused on the latter issue, dealing with the effects of microaggregation on the estimation of the simple linear regression model.

The main results concerning microaggregation with respect to the dependent variable Y are:

1. The naive least squares estimate $\tilde{\beta}$ is biased if the data are aggregated with respect to Y. The effect of the covariate X on the response Y is overestimated on average. The only exception is the case $\beta = 0$, where the naive least squares estimator yields a consistent estimate of the true β. If $|\beta| \to \infty$, $\tilde{\beta}$ becomes again asymptotically unbiased in the limit. By increasing the aggregation level A, bias($\tilde{\beta}$) increases as well and becomes more and more severe.
2. The above result shows that there is a major difference between aggregating the data with respect to Y and aggregating the data with respect to X. In the latter case, $\tilde{\beta}$ is unbiased for any value of β. Although aggregating with respect to X therefore seems to be more convenient for statistical analysis, it has to be pointed out that scientists do not necessarily know *in advance* which variable is the dependent variable Y.
3. The naive least squares estimates $\tilde{\alpha}$ and $\tilde{\sigma}^2_\epsilon$ show similar biases. Again, $\tilde{\alpha}$ and $\tilde{\sigma}^2_\epsilon$ are (asymptotically) unbiased for $\beta = 0$.
4. The asymptotic bias is a very good approximation to the finite sample bias even if the sample size is rather small.
5. It is possible to remove the bias and to construct consistent estimates of α, β, and σ^2_ϵ by correcting the naive least squares estimates (see section 4.4).
6. The corrected estimators show some bias for small samples (e.g., $n = 50$). For $\tilde{\beta}_c$ the bias is not very large, but for $\tilde{\sigma}^2_{\epsilon,c}$ the bias can be aggravating, although it becomes negligible again when $n > 150$.

In summary, the above results suggest that it is not advisable to apply standard linear model techniques to a microaggregated data set if the response variable Y has been used to determine the similarity of the data values. However, we have shown how to correct for the bias of $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\sigma}_\epsilon^2$ in order to get consistent estimates. In a subsequent paper, we will focus on the variances of $\tilde{\beta}$ and $\tilde{\beta}_c$. Moreover, the impact of microaggregation on the power of t-tests will be analyzed.

Acknowledgements

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

Appendix - Proofs

A Microaggregation with Respect to the Regressor

In the following, we use the notation of section 2. The data values are assumed to be aggregated with respect to the regressor X of model (1). Denote by $X_{sort(x)}$ the design matrix $(\mathbf{1}, x_{sort(x)})$.

Theorem 1. *The estimate $\tilde{\gamma}$ based on the microaggregated data is a consistent estimate of γ.*

Proof: Provided that $X'_{sort(x)}DX_{sort(x)}$ is nonsingular, we have

$$\begin{aligned}
\tilde{\gamma} &= (X'_{sort(x)}DX_{sort(x)})^{-1}X'_{sort(x)}Dy_{sort(x)} \\
&= (X'_{sort(x)}DX_{sort(x)})^{-1}X'_{sort(x)}D(X'_{sort(x)}\gamma + e_{sort(x)}) \\
&= \gamma + (X'_{sort(x)}DX_{sort(x)})^{-1}X'_{sort(x)}De_{sort(x)} \ , \qquad (22)
\end{aligned}$$

where $e_{sort(x)}$ denotes the error vector after sorting the data with respect to X. As $\mathrm{E}(e_{sort(x)}) = 0$, it follows that $\mathrm{E}(\tilde{\gamma}) = \gamma$. The variance of $\tilde{\gamma}$ becomes

$$\begin{aligned}
\mathrm{var}(\tilde{\gamma}) &= \mathrm{E}\big[(\tilde{\gamma}-\gamma)(\tilde{\gamma}-\gamma)'\big] \\
&= \mathrm{E}\big[(X'_{sort(x)}DX_{sort(x)})^{-1}X'_{sort(x)}De_{sort(x)} \\
&\quad \cdot e'_{sort(x)}DX_{sort(x)}(X'_{sort(x)}DX_{sort(x)})^{-1}\big] \\
&= \sigma^2_\epsilon\ (X'_{sort(x)}DX_{sort(x)})^{-1} \\
&= \frac{1}{n}\,\sigma^2_\epsilon\ (\frac{1}{n}X'_{sort(x)}DX_{sort(x)})^{-1} \ . \qquad (23)
\end{aligned}$$

If the variable X follows a distribution with variance σ_x^2, $\frac{1}{n}X'_{sort(x)}DX_{sort(x)}$ converges to a finite matrix (see Lemma 1(a) and the note at the end of its proof in appendix B). Therefore, var($\tilde{\gamma}$) converges to zero. It follows that $\tilde{\gamma}$ is a consistent estimate of γ.

For the proof of Theorem 2 we need the following lemma:

Lemma A. *Denote by $\tilde{e}'_x\tilde{e}_x$ the residual sum of squares based on the aggregated data. Then,* $\mathrm{E}(\tilde{e}'_x\tilde{e}_x) = (n/A-2)\sigma_\epsilon^2$, *provided that $X'_{sort(x)}DX_{sort(x)}$ is nonsingular.*

Proof: The residual sum of squares can be written as

$$\begin{aligned}\tilde{e}'_x\tilde{e}_x &= (\tilde{y}_x - \tilde{X}_x\tilde{\gamma})'(\tilde{y}_x - \tilde{X}_x\tilde{\gamma}) \\ &= y'_{sort(x)}\big(D - DX_{sort(x)}(X'_{sort(x)}DX_{sort(x)})^{-1}X'_{sort(x)}D\big)y_{sort(x)}\,. \qquad (24)\end{aligned}$$

It is easily seen that $Q := D - DX_{sort(x)}(X'_{sort(x)}DX_{sort(x)})^{-1}X'_{sort(x)}D$ is an idempotent matrix. Moreover, $QX_{sort(x)} = 0$, and thus

$$\begin{aligned}\tilde{e}'_x\tilde{e}_x &= y'_{sort(x)}Qy_{sort(x)} \\ &= (X_{sort(x)}\gamma + e_{sort(x)})'Q(X_{sort(x)}\gamma + e_{sort(x)}) \\ &= e'_{sort(x)}Qe_{sort(x)} \\ &= \mathrm{tr}\big[Q\ (e_{sort(x)}e'_{sort(x)})\big]\ . \qquad (25)\end{aligned}$$

Taking expectations we receive

$$\begin{aligned} \mathrm{E}(\tilde{e}_x'\tilde{e}_x) &= \sigma_\epsilon^2 \,\mathrm{tr}(Q) \\ &= \sigma_\epsilon^2 \left(\frac{n}{A} - 2\right) . \end{aligned} \tag{26}$$

Theorem 2. *$(A/n)\ \tilde{e}_x'\tilde{e}_x$ is a consistent estimator of σ_ϵ^2.*

Proof: From (25),

$$\begin{aligned} \mathrm{var}\left(\frac{1}{n}\,\tilde{e}_x'\tilde{e}_x\right) &= \frac{1}{n^2}\,\mathrm{var}(e_{sort(x)}'Qe_{sort(x)}) \\ &= \frac{\sigma_\epsilon^4}{n^2}\,\mathrm{var}\big((e_{sort(x)}/\sigma_\epsilon)'Q(e_{sort(x)}/\sigma_\epsilon)\big) . \end{aligned} \tag{27}$$

Now, $(e_{sort(x)}/\sigma_\epsilon)'\,Q\,(e_{sort(x)}/\sigma_\epsilon)$ follows a $\chi^2_{n/A-2}$-distribution, which implies

$$\mathrm{var}\left(\frac{1}{n}\,\tilde{e}_x'\tilde{e}_x\right) = \,2\left(\frac{n}{A} - 2\right)\,\frac{\sigma_\epsilon^4}{n^2} . \tag{28}$$

Therefore, $\mathrm{var}\big((A/n)\ \tilde{e}_x'\tilde{e}_x\big) \to 0$ as $n \to \infty$. On the other hand, (26) implies

$$\lim_{n\to\infty} \mathrm{E}\left(\frac{A}{n}\,\tilde{e}_x'\tilde{e}_x\right) = \sigma_\epsilon^2 . \tag{29}$$

This, together with (28), proves the theorem.

B Microaggregation with Respect to the Dependent Variable

In the following, we use the notation of section 4.

Lemma 1. *Denote by $S^2_{\tilde{y}_y}$ the empirical variance of $\tilde{y}_y$. Then the following results hold:*

a) $S^2_{\tilde{y}_y}$ *converges to* σ_y^2 *in probability.*

b) $S^2_{\tilde{x}_y}$ *converges in probability to* $\tilde{\sigma}_x^2 := \sigma_x^2 / f(\rho)$*, where*

$$f(\rho) := \frac{1}{\frac{1}{A} + \left(1 - \frac{1}{A}\right)\rho^2} \ . \tag{30}$$

c) $S_{\tilde{x}_y \tilde{y}_y}$ *converges in probability to* $\sigma_{xy} := \rho \sigma_x \sigma_y$*.*

Proof of a): Assume that the elements of y have been ordered according to their magnitude: $y_1 < y_2 < \cdots < y_n$, and that the data have been grouped into groups $G_i := \{y_{iA+1}, \ldots, y_{iA+A}\}$, $i = 0, \ldots, (n-A)/A$. Denote by $S^2_{y,W}$ and $S^2_{y,B}$ the within-groups and between-groups variances, respectively. By definition, $S^2_{y,B} = S^2_{\tilde{y}_y}$. As $S^2_{y,B} + S^2_{y,W}$ is equal to the empirical variance S_y^2 of y and as $S_y^2 \to \sigma_y^2$, we only have to show that $S^2_{y,W} \to 0$.

For any $\epsilon > 0$ let B be such that $\int_{|y|>B} y^2 \, dF(y) < \epsilon$ (this is possible because σ_y^2 exists). Now, $S^2_{y,W}$ can be written in the following way:

$$S^2_{y,W} = \frac{1}{n} \sum_{i=1}^{n/A} S_i^2 \ , \tag{31}$$

where $S_i^2 := \sum_{j \in G_i} (y_j - \bar{y}_i)^2$. Let u be such that $\min(G_u) < -B$, $\max(G_u) > -B$. In the same way, define o such that $\min(G_o) < B$, $\max(G_o) > B$ (compare Fig. 17).

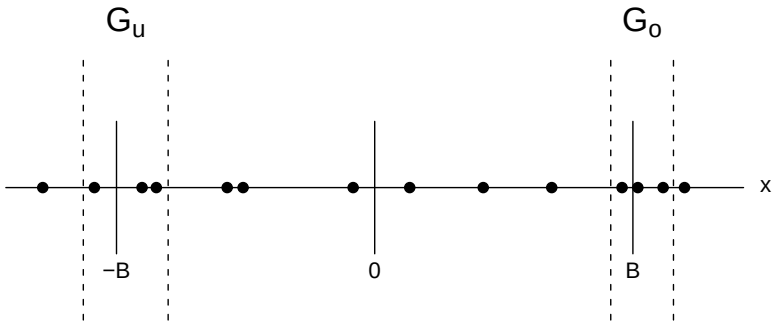

Figure 17: Microaggregation with respect to Y - definition of groups

Furthermore, let $\bar{G}_i := \{|y_{iA+1}|, \ldots, |y_{iA+A}|\}$. By defining

$$S_{y,W1}^2 := \frac{1}{n} \sum_{i:\max(\bar{G}_i)<B} S_i^2 \ , \tag{32}$$

$$S_{y,W2}^2 := \frac{1}{n} \sum_{i:\min(\bar{G}_i)>B} S_i^2 \ , \tag{33}$$

$$S_{y,W3}^2 := \frac{1}{n} \left(S_u^2 + S_o^2\right) \ , \tag{34}$$

$S_{y,W}^2$ can be subdivided into

$$S_{y,W}^2 = S_{y,W1}^2 + S_{y,W2}^2 + S_{y,W3}^2 \ . \tag{35}$$

Clearly, $S^2_{y,W3} \to 0$ if $n \to \infty$. As

$$\begin{aligned} S^2_{y,W2} &\leq \frac{1}{n} \sum_{|y_j|>B} y_j^2 \\ &\to \int_{|y|>B} y^2 \, dF(y) < \epsilon \ , \end{aligned} \tag{36}$$

$S^2_{y,W2} \to 0$, too.

Finally, it can be shown that $S^2_{y,W1} \to 0$: Divide $[-B, B]$ into intervals of length $\sqrt{\epsilon}/A$. If each interval contains at least one observation y_j, then $S^2_{y,W1} \leq \frac{1}{n} \frac{n}{A} A (A\frac{\sqrt{\epsilon}}{A})^2 = \epsilon$. As will be seen, the probability of this event goes to one.

Denote by A_k the event that at least one observation y_j, $j = 1, \ldots, n$ lies in the kth interval I_k. Then $\mathrm{P}(\bar{A}_k) = (1 - q_k)^n$, where $q_k := \mathrm{P}(Y \in I_k)$. Because Y is a continuous variable, $q_k > 0$. Therefore, $\lim_{n\to\infty} \mathrm{P}(\bar{A}_k) = 0$ for each k. It follows that

$$\mathrm{P}\Big(\bigcap_k A_k\Big) \ = \ 1 - \mathrm{P}\Big(\bigcup_k \bar{A}_k\Big) \ \geq \ 1 - \sum_k \mathrm{P}(\bar{A}_k) \ \stackrel{n\to\infty}{\longrightarrow} \ 1 \ , \tag{37}$$

and thus $\mathrm{P}\big(S^2_{y,W1} \leq \epsilon\big)$ converges to one as well.

Note that the only assumption we needed to prove Lemma 1a) was the existence of σ_y^2. Normality of Y was not required.

Proof of b): To show Lemma 1b), we use the theory of induced order statistics (see, e.g., David(1981)). Suppose we have an iid sample (x_i, y_i), $i = 1, \ldots, n$ from a bivariate normal distribution with variances σ_x^2, σ_y^2 and correlation ρ. Without

loss of generality we set $\mu_x = \mu_y = 0$. Denote by $Y_{r:n}$ the r-th order statistic of Y. The X-variable associated with $Y_{r:n}$ is then called *induced order statistic* or *concomitant* $X_{[r:n]}$.

As X and Y are jointly normally distributed, we have

$$X_i = \beta^* Y_i + \delta_i \ , \ i = 1, \ldots, n, \tag{38}$$

where $\mathrm{E}(\delta_i) = 0$ and δ_i is independent of Y_i. It follows that

$$X_{[r:n]} = \beta^* Y_{r:n} + \delta_{[r]} \ , \tag{39}$$

where $\delta_{[r]}$ denotes the random variable associated with $Y_{r:n}$. $Y_{r:n}$ and $\delta_{[r]}$ are independent. Moreover, $\delta_{[1]}, \ldots, \delta_{[n]}$ are independent and identically distributed with zero mean and variance $\sigma_\delta^2 = (1-\rho^2)\sigma_x^2$. Now, with $\tilde{\delta}_y$ denoting the vector containing the aggregated values of δ, we have

$$\tilde{x}_y = \beta^* \tilde{y}_y + \tilde{\delta}_y \ , \tag{40}$$

where $\tilde{y}_{y,i}$ and $\tilde{\delta}_{y,i}$, $i = 1, \ldots, n$, are independent. By (40), we have

$$S^2_{\tilde{x}_y} = \beta^{*2} S^2_{\tilde{y}_y} + S^2_{\tilde{\delta}_y} + 2\beta^* S_{\tilde{y}_y \tilde{\delta}_y} \ , \tag{41}$$

where $S^2_{\tilde{\delta}_y}$ denotes the empirical variance of $\tilde{\delta}_y$ and $S_{\tilde{y}_y \tilde{\delta}_y}$ denotes the empirical covariance of $\tilde{y}_y$ and $\tilde{\delta}_y$.

As $S^2_{\tilde{y}_y} \to \sigma_y^2$, see Lemma 1a), $S^2_{\tilde{\delta}_y} \to (1/A)\,\sigma_\delta^2 = (1/A)\,(1-\rho^2)\,\sigma_x^2$, and $S_{\tilde{y}_y\tilde{\delta}_y} \to \sigma_{y\delta} = 0$, we have, with $\beta^* = \rho\sigma_x/\sigma_y$,

$$\begin{aligned} S^2_{\tilde{x}_y} &\to \beta^{*2}\sigma_y^2 + \frac{1}{A}\,\sigma_x^2(1-\rho^2) \\ &= \sigma_x^2\,\big(\frac{1}{A} + \big(1-\frac{1}{A}\big)\rho^2\big) \\ &= \frac{\sigma_x^2}{f(\rho)}. \end{aligned} \tag{42}$$

Proof of c): From Lemma 1a) we know that $S^2_{\tilde{y}_y} \to \sigma_y^2$. Moreover, Theorem 1 yields

$$\frac{S_{\tilde{x}_y\tilde{y}_y}}{S^2_{\tilde{y}_y}} \to \beta^* = \frac{\sigma_{xy}}{\sigma_y^2}\ . \tag{43}$$

Hence the lemma is proved. Note that $S_{\tilde{x}_x\tilde{y}_x}$ and $S_{\tilde{x}_y\tilde{y}_y}$ do not only have the same limit σ_{xy}. They also have the same mean (see Lemma B).

Lemma B. *Denote by $S_{\tilde{x}_x\tilde{y}_x}$ the empirical covariance of $\tilde{x}_x$ and $\tilde{y}_x$. Then,* $\mathrm{E}(S_{\tilde{x}_x\tilde{y}_x}) = \mathrm{E}(S_{\tilde{x}_y\tilde{y}_y})$.

Proof: Let us first assume that X and Y have equal variances $\sigma_x^2 = \sigma_y^2 =: \sigma^2$. As X and Y are jointly normally distributed random variables with density f, we have $f(x,y) = f(y,x)$. Define

$$c_1(x,y) := S_{\tilde{x}_x\tilde{y}_x} \tag{44}$$

$$c_2(x,y) := S_{\tilde{x}_y\tilde{y}_y}\ . \tag{45}$$

As

$$c_2(y,x) = S_{\tilde{y}_x\tilde{x}_x} = S_{\tilde{x}_x\tilde{y}_x} = c_1(x,y)\ , \tag{46}$$

it follows that

$$\begin{aligned} \mathrm{E}(S_{\tilde{x}_y\tilde{y}_y}) &= \int c_2(x,y)f(x,y)\ d(x,y) \\ &= \int c_2(y,x)f(y,x)\ d(y,x) \\ &= \int c_1(x,y)f(x,y)\ d(x,y) \\ &= \mathrm{E}(S_{\tilde{x}_x\tilde{y}_x}) \end{aligned} \tag{47}$$

If X and Y have nonequal variances σ_x^2 and σ_y^2, equation (47) still holds. This is because the scaling of X and Y does not affect the ordering of (x_i, y_i), $i = 1, \ldots, n$. Therefore,

$$\mathrm{S}_{\tilde{x}_x\tilde{y}_x} = \sigma_x\sigma_y\ \mathrm{S}_{\tilde{x}^*_{x^*}\tilde{y}^*_{x^*}} \quad \text{and} \tag{48}$$

$$\mathrm{S}_{\tilde{x}_y\tilde{y}_y} = \sigma_x\sigma_y\ \mathrm{S}_{\tilde{x}^*_{y^*}\tilde{y}^*_{y^*}}\ , \tag{49}$$

where X^* and Y^* denote the standardized variables corresponding to X and Y. Clearly, X^* and Y^* have equal variances $\sigma^2_{x^*} = \sigma^2_{y^*} = 1$. It follows from (47) that

$$\begin{aligned} \mathrm{E}(\mathrm{S}_{\tilde{x}_x\tilde{y}_x}) &= \sigma_x\sigma_y\ \mathrm{E}(\mathrm{S}_{\tilde{x}^*_{x^*}\tilde{y}^*_{x^*}}) \\ &= \sigma_x\sigma_y\ \mathrm{E}(\mathrm{S}_{\tilde{x}^*_{y^*}\tilde{y}^*_{y^*}}) \\ &= \mathrm{E}(\mathrm{S}_{\tilde{x}_y\tilde{y}_y})\ . \end{aligned} \tag{50}$$

References

[1] Anwar, M. N. (1993). *Micro-Aggregation - The Small Aggregates Method*. Internal report, Luxemburg, Eurostat.

[2] Brand, R. (2000). *Anonymität von Betriebsdaten*. Beiträge zur Arbeitsmarkt- und Berufsforschung BeitrAB 237, Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

[3] David, H. A. (1981). *Order Statistics, Second Edition*. Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

[4] 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.

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

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

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

[8] Gottschalk, S. (2004). *Klassische Anonymisierung und Resampling von Unternehmensdaten - Trade-off zwischen Datenschutz und Analysepotenzial*. Dissertation, Fakultät für Wirtschaftswissenschaften, Universität Bielefeld.

[9] Köhler, S. (1999). *Anonymisierung von Mikrodaten in der Bundesrepublik und ihre Nutzung - ein Überblick*. In *Methoden zur Sicherung der statistischen Geheimhaltung*, Forum der Bundesstatistik, Band 31, Wiesbaden, 133-149.

[10] Lechner, S. and Pohlmeier, W. (2003). *Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten*. In *Anonymisierung wirtschaftsstatistischer Einzeldaten* (G. Ronning, R. Gnoss, eds.), Forum der Bundesstatistik, Band 42, Wiesbaden, 115-137.

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

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

[13] Rosemann, M. (2004). *Auswirkungen unterschiedlicher Varianten der Mikroaggregation auf die Ergebnisse linearer und nicht linearer Schätzungen*. Contribution to the Workshop *Econometric Analysis of Anonymized Firm Data*, Tübingen, Institut für Angewandte Wirtschaftsforschung, March 18-19, 2004.