Schmid, Matthias; Schneeweiss, Hans; Küchenhoff, Helmut

Working Paper

Statistical inference in a simple linear model under microaggregation

Discussion Paper, No. 416

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) : Statistical inference in a simple linear model under microaggregation, Discussion Paper, No. 416, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München,
https://doi.org/10.5282/ubm/epub.1785

This Version is available at:
https://hdl.handle.net/10419/31144

Statistical Inference in 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 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. We focus on the effect of microaggregation on a simple linear model. In a previous paper we have shown how to correct for the aggregation bias of the naive least-squares estimator that occurs when the dependent variable is used to group the data. The present paper deals with the asymptotic variance of the corrected least-squares estimator and with the asymptotic variance of the naive least-squares estimator when either the dependent variable or the regressor is used to group the data. We derive asymptotic confidence intervals for the slope parameter. Furthermore, we show how to test for the significance of the slope parameter by analyzing the effect of microaggregation on the asymptotic power function of the naive t-test.

Keywords: Microaggregation, simple linear model, asymptotic variance, t-test, disclosure control

1 Introduction

The development of empirical research as well as the growing capacity of modern computer systems have led to an increasing demand on microdata over the last decades. Statistical offices and other data providers are therefore faced with the problem of providing sufficient information to scientists while at the same time having to maintain confidentiality required by data protection laws. One method to handle this trade-off (which is commonly referred to as the *statistical disclosure control problem*) is the dissemination of factually anonymized data sets, also called scientific-use files. The idea behind the creation of scientific-use files is the reduction of the information content of a data set by means of masking procedures. However, while reducing the disclosure risk of a data file, masking procedures also affect the results of statistical analyses.

One of the most promising masking techniques is microaggregation, a procedure for continuous data which has been widely discussed over the last years (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. In the literature, many suggestions have been made on how to form the groups (see, e.g., Domingo-Ferrer and Mateo-Sanz (2002)). To reduce the information loss imposed by microaggregation, it is considered advisable to group only those data values which are similar in terms of a similarity criterion.

In Schmid, Schneeweiss and Küchenhoff (2005) we have studied a microaggregation technique that uses a so-called "leading variable" to form the groups (Paass and Wauschkuhn (1985), Mateo-Sanz and Domingo-Ferrer (1998)). This procedure subdivides the data set into groups having similar values for the leading variable. We have analyzed the effects of this kind of microaggregation on the estimation of a simple linear regression model. Interestingly, the properties of the resulting linear model estimates depend on the choice of the leading variable: If the regressor X serves as the leading variable, estimates are unbiased although having greater variance (see also Feige and Watts (1972) or Lechner and Pohlmeier (2003)). If the dependent variable Y serves as the leading variable, estimates are biased. However, the bias can be removed and consistent estimators for the slope parameter, the intercept and the residual error variance of the model can be constructed.

This paper is a continuation of Schmid et al. (2005). Again, we consider the estimation of a simple linear regression model with microaggregated data. The focus now is on testing and the construction of confidence intervals for the slope parameter β. By means of the delta method, formulas for the variances of the naive least squares estimators and the corrected least squares estimator of β are derived. Thus, an asymptotic confidence interval for the slope parameter can be constructed. Moreover, to assess whether β is significantly different from zero, we construct a t-test which asymptotically has the same power function as the t-test based on the original data. In addition to the theoretical results, we carry out a systematic simulation study to examine the small sample properties of our proposed procedures.

In section 2, we briefly summarize the results presented in Schmid et al. (2005).

Section 3 deals with the asymptotic variances of the naive and the corrected least squares estimators of the slope parameter β. In section 4 we show how to carry out t-tests with microaggregated data. In section 5, a systematic simulation study on the results derived in sections 3 and 4 is carried out. Section 6 contains a concluding summary. Proofs are relegated to the appendix.

2 Consistent Estimation of a Simple Linear Model with Microaggregated Data

In this section, the results of Schmid et al. (2005) are briefly summarized. We consider the simple linear model

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

Y denotes the continuous response (or endogenous variable) while X denotes the continuous covariate (or exogenous variable). $\gamma := (\alpha, \beta)'$ 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. In the following, we use a fixed group size (also called aggregation level) A. As stated in the introduction, the data can either be aggregated with respect to the leading variable X or with respect to the leading variable Y. In both cases, microaggregation works as follows: First, the data vectors x and y have to be sorted with respect to

the leading variable. The sorted data set is then 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 are averaged and the averages are assigned to the items of the group.

Denote by $\tilde{y}_x$ and $\tilde{x}_x$ the vectors containing the data that have been aggregated with respect X. Similarly, denote by $\tilde{y}_y$ and $\tilde{x}_y$ the data vectors if microaggregation with respect to Y has been performed. Further, denote the empirical variance of $\tilde{x}_x$ and $\tilde{x}_y$ computed from the microaggregated data by $S^2_{\tilde{x}_x} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_{x,i} - \bar{\tilde{x}}_x)^2$ and $S^2_{\tilde{x}_y} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_{y,i} - \bar{\tilde{x}}_y)^2$, respectively. The variances of $\tilde{y}_x$ and $\tilde{y}_y$ and the covariances of $\tilde{x}_x$ and $\tilde{y}_x$ and of $\tilde{x}_y$ and $\tilde{y}_y$ are denoted in a similar way, e.g. $S_{\tilde{x}_x\tilde{y}_x} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_{x,i} - \bar{\tilde{x}}_x)(\tilde{y}_{x,i} - \bar{\tilde{y}}_x)$.

Now, if the data are microaggregated with respect to X and the slope parameter β is estimated by ordinary least squares, i.e.

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

then $\tilde{\beta}_x$ is an unbiased and consistent estimator of β. (The same can be said of the naive least squares estimator $\tilde{\alpha}_x$).

If Y is used as the leading variable, we need the additional assumption that X follows a normal distribution with mean μ_x and variance σ_x^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 $\sigma_y^2 := \beta^2\sigma_x^2 + \sigma_\epsilon^2$. The OLS estimator $\tilde{\beta}_y = \frac{S_{\tilde{x}_y\tilde{y}_y}}{S^2_{\tilde{x}_y}}$ then

converges in probability to $f(\rho)\beta$, where

$$f(\rho) := \frac{1}{\frac{1}{A} + \left(1 - \frac{1}{A}\right)\rho^2} \tag{3}$$

and ρ is the correlation of X and Y.

With the help of (3), it is possible to derive a consistent estimator of $\tilde{\beta}_y$: Denote by $\tilde{\rho}_y$ the empirical correlation coefficient between $\tilde{x}_y$ and $\tilde{y}_y$. A consistent estimate of ρ can be obtained from

$$\tilde{\rho}^2_{y,c} := \frac{\tilde{\rho}^2_y}{A - (A-1)\tilde{\rho}^2_y} \ . \tag{4}$$

The corrected estimator of $\tilde{\beta}_y$ then becomes

$$\tilde{\beta}_{y,c} = \frac{\tilde{\beta}_y}{A - (A-1)\tilde{\rho}^2_y} \ . \tag{5}$$

While consistent estimation of the parameters in model (1) is crucial, it is equally important to derive variance formulas of the estimators in order to calculate confidence intervals. This will be the subject of the next section.

3 Variances of $\tilde{\beta}_x$, $\tilde{\beta}_y$, and $\tilde{\beta}_{y,c}$

In the following, we derive the asymptotic variances of $\tilde{\beta}_x$, $\tilde{\beta}_y$, and $\tilde{\beta}_{y,c}$. To achieve this, some additional notation is required first:

- Two random sequences a_n and b_n are said to be asymptotically *equivalent* if $\text{plim}_{n\to\infty}\sqrt{n}(a_n - b_n) = 0$. We write $a_n \sim b_n$.

- They are said to be asymptotically *equal* if $\text{plim}_{n\to\infty}(a_n - b_n) = 0$. We write $a_n \approx b_n$. Thus $a_n \sim b_n$ is the same as $\sqrt{n}a_n \approx \sqrt{n}b_n$.

Moreover, we say for short "the asymptotic variance of a random sequence a_n is equal to σ_a^2/n" if $\text{plim}_{n\to\infty}a_n =: \alpha$ exists and if $\sqrt{n}(a_n - \alpha)$ converges in distribution to $\text{N}(0, \sigma_a^2)$ as $n \to \infty$. The asymptotic variance of a_n is then denoted by $\text{var}(a_n) = \sigma_a^2/n$.

3.1 Asymptotic Properties of the Naive Variance Estimates

In the sequel, we will make use of the following fundamental lemma, which compares the empirical variances and covariances of the aggregated variables to those of the original, non-aggregated variables X and Y. We formulate the lemma in terms of aggregation with respect to Y. By interchanging the role of X and Y, a corresponding lemma can be stated in terms of aggregation with respect to X.

We will assume throughout that X and Y are jointly normally distributed with parameters μ_x, μ_y, σ_x^2, σ_y^2, and $\sigma_{xy} := \rho\sigma_x\sigma_y$. In addition, we will make use of the regression model

$$X = \alpha^* + \beta^* Y + \delta \ , \tag{6}$$

where β^* is equal to σ_{xy}/σ_y^2. The error variable δ has mean zero and variance $\sigma_\delta^2 := (1-\rho^2)\sigma_x^2$. Moreover, as X and Y are jointly normally distributed, Y and δ are independent. Denote by S_δ^2 the empirical variance of the (unobserved) values $\delta_1, \ldots, \delta_n$ and denote by $S_{\tilde{\delta}_y}^2$ the empirical variance of the aggregated values $(\tilde{\delta}_{y,1}, \ldots, \tilde{\delta}_{y,n})' =: \tilde{\delta}_y$.

Lemma 1.

a) Denote by S_y^2 *the empirical variance of* y*. Then* $\sqrt{n}(S_{\tilde{y}_y}^2 - S_y^2)$ *converges in probability to 0.*

b) Denote by $S_{y\delta}$ *the empirical covariance of* Y *and* δ *in model (6). Analogously, denote by* $S_{\tilde{y}_y\tilde{\delta}_y}$ *the empirical covariance of* $\tilde{y}_y$ *and* $\tilde{\delta}_y$ *Then,* $\sqrt{n}(S_{\tilde{y}_y\tilde{\delta}_y} - S_{y\delta})$ *converges in probability to 0.*

c) Denote by S_{xy} *the empirical covariance of* x *and* y*. Then* $\sqrt{n}(S_{\tilde{x}_y\tilde{y}_y} - S_{xy})$ *converges in probability to 0.*

d) Denote by S_x^2 *the empirical variance of* x*. Then* $S_{\tilde{x}_y}^2 - S_x^2$ *is asymptotically equivalent to* $S_{\tilde{\delta}_y}^2 - S_\delta^2$*.*

e) For $n \to \infty$*,* $\sqrt{n}(S_{\tilde{\delta}_y}^2 - \frac{1}{A}S_\delta^2)$ *converges to a normal distribution with zero mean and variance* $2\frac{A-1}{A^2}\sigma_\delta^4 = 2\frac{A-1}{A^2}\sigma_x^4(1-\rho^2)^2$*.*

f) Consider the equation

$$X_i = \hat{\beta}^* Y_i + \hat{\delta}_i \ , \ i = 1, \ldots, n \ , \tag{7}$$

where $\hat{\beta}^*$ *is the least squares estimate based on the non-aggregated data and* $\hat{\delta}_i := X_i - \hat{\beta}^* Y_i$ *is the corresponding residual term. Then,* S_δ^2 *is asymptotically equivalent to the empirical variance* $S_{\hat{\delta}}^2$ *of* $\hat{\delta}_1, \ldots, \hat{\delta}_n$*.*

Proof: See appendix.

3.2 Variance of $\tilde{\beta}_x$

To derive the asymptotic variance of $\tilde{\beta}_x$, we make use of the following theorem:

Theorem 1. *$\tilde{\beta}_x$ is asymptotically equivalent to $\hat{\beta}$, where $\hat{\beta}$ is the least squares estimate of β computed from the original (non-aggregated) data. Consequently, $\sqrt{n}(\tilde{\beta}_x - \beta) \xrightarrow{d} N(0, v^2)$, where $v^2 := \sigma_\epsilon^2/\sigma_x^2$.*

Proof: By definition of $\tilde{\beta}_x$ and $\hat{\beta}$, we have

$$\begin{aligned}\sqrt{n}(\tilde{\beta}_x - \hat{\beta}) &= \sqrt{n}\Big(\frac{S_{\tilde{x}_x\tilde{y}_x}}{S^2_{\tilde{x}_x}} - \frac{S_{xy}}{S^2_x}\Big) \\ &= \frac{1}{S^2_{\tilde{x}_x}}\sqrt{n}(S_{\tilde{x}_x\tilde{y}_x} - S_{xy}) - \frac{S_{xy}}{S^2_x S^2_{\tilde{x}_x}}\sqrt{n}(S^2_{\tilde{x}_x} - S^2_x)\ . \qquad (8)\end{aligned}$$

By Lemma 1, with the roles of x and y interchanged, (8) goes to zero as $n \to \infty$.

Thus, we have

$$\mathrm{var}(\tilde{\beta}_x) = \frac{v^2}{n}\ . \qquad (9)$$

Note that this is the same asymptotic variance as the one for $\hat{\beta}$. $\tilde{\beta}_x$ and $\hat{\beta}$ are asymptotically equally efficient. Define $S^2_{\tilde{e}_x} := S^2_{\tilde{y}_x} - \tilde{\beta}^2_x S^2_{\tilde{x}_x}$. In Schmid et al. (2005), we have shown that $A \cdot S^2_{\tilde{e}_x}/S^2_{\tilde{x}_x}$ is a consistent estimator of v^2. An asymptotic confidence interval for β is thus given by

$$\left[\ \tilde{\beta}_x - z_{1-\alpha/2}\sqrt{\frac{AS^2_{\tilde{e}_x}}{nS^2_{\tilde{x}_x}}}\ ,\ \tilde{\beta}_x + z_{1-\alpha/2}\sqrt{\frac{AS^2_{\tilde{e}_x}}{nS^2_{\tilde{x}_x}}}\ \right]\ , \qquad (10)$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

3.3 Variance of $\tilde{\beta}_y$

As explained above, $S_{\tilde{x}_y\tilde{y}_y}$ denotes the empirical covariance of $\tilde{x}_y$ and $\tilde{y}_y$. To derive the asymptotic variance of $\tilde{\beta}_y$, we express $\tilde{\beta}_y$ as a function of $\tilde{S} := (S^2_{\tilde{x}_y}, S^2_{\tilde{y}_y}, S_{\tilde{x}_y\tilde{y}_y})'$:

$$\tilde{\beta}_y = \frac{S_{\tilde{x}_y\tilde{y}_y}}{S^2_{\tilde{x}_y}} =: F(\tilde{S}) \ . \tag{11}$$

Note that F does not depend on $S^2_{\tilde{y}_y}$. However, $S^2_{\tilde{y}_y}$ will be needed later to derive the asymptotic variance of $\tilde{\beta}_{y,c}$.

Now, if a formula for the asymptotic covariance matrix $\mathrm{cov}(\tilde{S})$ of $(\tilde{S} - \mathrm{plim}\tilde{S})$ is found, we can obtain the asymptotic variance of $\tilde{\beta}_y$ by applying the delta method.

In Schmid et al. (2005) we proved that

$$\mathrm{plim}\,\tilde{S} = \begin{pmatrix} \sigma_x^2/f(\rho) \\ \sigma_y^2 \\ \sigma_{xy} \end{pmatrix} =: \bar{S} \ . \tag{12}$$

Denote by $\hat{\rho}$ the empirical correlation coefficient based on the non-aggregated data. Then, $(\tilde{S} - \mathrm{plim}\tilde{S})$ can be reduced to expressions in S_x^2, S_y^2, and $S_{xy} = \hat{\rho}\sqrt{S_x^2 S_y^2}$ plus an independent term of known variance. To achieve this, we again make use of the regression model (6).

From Lemma 1, we can derive the following result:

Lemma 2. *$S^2_{\tilde{x}_y}$ is asymptotically equivalent to $\frac{S_x^2}{f(\rho)} + (1 - \frac{1}{A})\sigma_x^2(\hat{\rho}^2 - \rho^2) + (S^2_{\tilde{\delta}_y} - \frac{1}{A}S^2_\delta)$.*

Proof: See appendix.

From Lemma 2 it follows that

$$\begin{aligned} S^2_{\tilde{x}_y} - \frac{\sigma^2_x}{f(\rho)} \quad &\sim \quad \frac{1}{f(\rho)}(S^2_x - \sigma^2_x) \\ &\quad + (1 - \frac{1}{A})\,\sigma^2_x\,(\hat{\rho}^2 - \rho^2) \\ &\quad + S^2_{\tilde{\delta}} - \frac{1}{A}S^2_\delta \ . \end{aligned} \tag{13}$$

Similarly, from Lemma 1, $(S^2_{\tilde{y}_y} - \sigma^2_y) \sim (S^2_y - \sigma^2_y)$ and $(S_{\tilde{x}_y\tilde{y}_y} - \sigma_{xy}) \sim (S_{xy} - \sigma_{xy})$.

Therefore, we have

$$\begin{aligned} \tilde{S} - \bar{S} \quad &\sim \quad \begin{pmatrix} \frac{1}{f(\rho)}(S^2_x - \sigma^2_x) + (1 - \frac{1}{A})\sigma^2_x(\hat{\rho}^2 - \rho^2) \\ S^2_y - \sigma^2_y \\ \hat{\rho}\sqrt{S^2_x S^2_y} - \sigma_{xy} \end{pmatrix} + \begin{pmatrix} S^2_{\tilde{\delta}} - \frac{1}{A}S^2_\delta \\ 0 \\ 0 \end{pmatrix} \\ &=: \quad G \begin{pmatrix} S^2_x \\ S^2_y \\ \hat{\rho} \end{pmatrix} + \begin{pmatrix} S^2_{\tilde{\delta}} - \frac{1}{A}S^2_\delta \\ 0 \\ 0 \end{pmatrix} \ . \end{aligned} \tag{14}$$

As the product moments $\mathrm{E}\big(S^2_x \cdot (S^2_{\tilde{\delta}} - \frac{1}{A}S^2_\delta)\big)$, $\mathrm{E}\big(S^2_y \cdot (S^2_{\tilde{\delta}} - \frac{1}{A}S^2_\delta)\big)$, and $\mathrm{E}\big(S_{xy} \cdot (S^2_{\tilde{\delta}} - \frac{1}{A}S^2_\delta)\big)$ are all equal to zero (compare equations (60) and (61) in the appendix), it can be shown that $S_1 := (S^2_x, S^2_y, \hat{\rho})'$ and $S_2 := \big((S^2_{\tilde{\delta}} - \frac{1}{A}S^2_\delta), 0, 0\big)'$ are asymptotically independent. Thus,

$$\mathrm{cov}(\tilde{S}) = \mathrm{cov}(\tilde{S} - \bar{S}) = \mathrm{cov}(G(S_1)) + \mathrm{cov}(S_2) \ . \tag{15}$$

Using (15), we can compute the asymptotic covariance of $\tilde{S}$ in the following way:

1. $\mathrm{cov}(G(S_1))$ can be evaluated by means of the delta method. Using the formulas derived in Kendall and Stuart (1977), we have

$$\Sigma_1 := \mathrm{cov}(S_1) = \frac{1}{n}\begin{pmatrix} 2\sigma_x^4 & 2\rho^2\sigma_x^2\sigma_y^2 & \sigma_x^2\rho(1-\rho^2) \\ 2\rho^2\sigma_x^2\sigma_y^2 & 2\sigma_y^4 & \sigma_y^2\rho(1-\rho^2) \\ \sigma_x^2\rho(1-\rho^2) & \sigma_y^2\rho(1-\rho^2) & (1-\rho^2)^2 \end{pmatrix} \tag{16}$$

and

$$D_1 := \begin{pmatrix} \frac{\partial G}{\partial S_x^2} & \frac{\partial G}{\partial S_y^2} & \frac{\partial G}{\partial \hat{\rho}} \end{pmatrix}\Big|_{(\sigma_x^2,\sigma_y^2,\rho)} = \begin{pmatrix} \frac{1}{f(\rho)} & 0 & 2(1-\frac{1}{A})\sigma_x^2\rho \\ 0 & 1 & 0 \\ \frac{1}{2}\frac{\sigma_y}{\sigma_x}\rho & \frac{1}{2}\frac{\sigma_x}{\sigma_y}\rho & \sigma_x\sigma_y \end{pmatrix} . \tag{17}$$

Therefore, $\mathrm{cov}(G(S_1)) = D_1\Sigma_1 D_1'$.

2. As $\mathrm{var}(S_{\tilde{\delta}}^2 - \frac{1}{A}S_{\delta}^2) = \frac{2}{n}\frac{A-1}{A^2}\sigma_x^4(1-\rho^2)^2$, see Lemma 1e), it follows that

$$\Sigma_2 := \mathrm{cov}(S_2) = \frac{1}{n}\begin{pmatrix} 2\frac{A-1}{A^2}\sigma_x^4(1-\rho^2)^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} . \tag{18}$$

We thus obtain

$$\mathrm{cov}(\tilde{S}) = D_1\Sigma_1 D_1' + \Sigma_2 . \tag{19}$$

Finally, with the help of (11) and (19), we can derive the asymptotic variance of $\tilde{\beta}_y$:

Theorem 2. *Define*

$$d := \frac{\partial F}{\partial \tilde{S}}\Big|_{(\sigma_x^2,\sigma_y^2,\rho)} = \begin{pmatrix} -\frac{\sigma_y}{\sigma_x^3}\rho f(\rho)^2 \\ 0 \\ f(\rho)/\sigma_x^2 \end{pmatrix} . \tag{20}$$

Then, $\mathrm{var}(\tilde{\beta}_y) = d'(D_1\Sigma_1 D_1' + \Sigma_2)d$.

Proof: By using the delta method, we obtain

$$\mathrm{var}(\tilde{\beta}_y) = \mathrm{var}(F(\tilde{S})) = d'\mathrm{cov}(\tilde{S})d = d'(D_1\Sigma_1 D_1' + \Sigma_2)d \ . \tag{21}$$

3.4 Variance of $\tilde{\beta}_{y,c}$

The asymptotic variance of the corrected estimator $\tilde{\beta}_{y,c}$ can be obtained in the same way as the asymptotic variance of the naive least squares estimator $\tilde{\beta}$. First of all, $\tilde{\beta}_{y,c}$ (see (5)) can be written as a function of $\tilde{S}$:

$$\begin{aligned} \tilde{\beta}_{y,c} &= \frac{S_{\tilde{x}_y\tilde{y}_y}/S^2_{\tilde{x}_y}}{A-(A-1)S^2_{\tilde{x}_y\tilde{y}_y}/(S^2_{\tilde{x}_y}S^2_{\tilde{y}_y})} \\ &= \frac{S_{\tilde{x}_y\tilde{y}_y}S^2_{\tilde{y}_y}}{AS^2_{\tilde{x}_y}S^2_{\tilde{y}_y}-(A-1)S^2_{\tilde{x}_y\tilde{y}_y}} =: F_c(\tilde{S}) \ . \end{aligned} \tag{22}$$

Defining

$$\tilde{N} := \left(\frac{A}{f(\rho)} - (A-1)\rho^2\right)^2 , \tag{23}$$

we obtain

$$d_c := \left.\frac{\partial F_c}{\partial \tilde{S}}\right|_{(\sigma_x^2,\sigma_y^2,\rho)} = \frac{1}{\tilde{N}} \begin{pmatrix} -\frac{A\rho\sigma_y}{\sigma_x^3} \\ -\frac{(A-1)\rho^3}{\sigma_x\sigma_y} \\ \frac{A+(A-1)f(\rho)\rho^2}{f(\rho)\sigma_x^2} \end{pmatrix} . \tag{24}$$

With the help of (15) and (24), we can derive $\mathrm{var}(\tilde{\beta}_{y,c})$:

Theorem 3. *The asymptotic variance of $\tilde{\beta}_{y,c}$ is equal to $d_c'(D_1\Sigma_1 D_1' + \Sigma_2)d_c$.*

Proof: By applying the delta method, we obtain

$$\mathrm{var}(\tilde{\beta}_{y,c}) = d_c' \mathrm{cov}(\tilde{S}) d_c = d_c'(D_1 \Sigma_1 D_1' + \Sigma_2) d_c \ . \tag{25}$$

In the special case where $\beta = \rho = 0$, it is easily seen that $\mathrm{var}(\tilde{\beta}_{y,c}) = v^2/n$, implying that $\tilde{\beta}_{y,c}$ and the estimator $\hat{\beta}$ based on the non-aggregated data are asymptotically equally efficient. Similarly, with some algebra, it can be shown that if $|\beta| \to \infty$, $\mathrm{var}(\tilde{\beta}_{y,c}) \to v^2/n$.

Substituting $\tilde{\rho}_{y,c}$ for ρ, $S^2_{\tilde{y}_y}$ for σ^2_y, and $f(\tilde{\rho}_{y,c}) S^2_{\tilde{x}_y}$ for σ^2_x, (25) can be consistently estimated. An asymptotic confidence interval for β is thus given by

$$\left[\ \tilde{\beta}_{y,c} - z_{1-\alpha/2} \sqrt{\tilde{\sigma}^2_{\tilde{\beta}_{y,c}}}, \tilde{\beta}_{y,c} + z_{1-\alpha/2} \sqrt{\tilde{\sigma}^2_{\tilde{\beta}_{y,c}}} \ \right] \ , \tag{26}$$

where $\tilde{\sigma}^2_{\tilde{\beta}_{y,c}}$ denotes the consistent estimate of $\mathrm{var}(\tilde{\beta}_{y,c})$.

4 T-Tests with Microaggregated Data

4.1 Microaggregation with Respect to X

In this section, the consequences of testing the null hypothesis "$H_0 : \beta = 0$" versus the alternative hypothesis "$H_1 : \beta \neq 0$" with microaggregated data are analyzed. Let us first consider the case where the data are aggregated with respect to X. An obvious approach is to assess the significance of the unbiased parameter estimate $\tilde{\beta}_x$ by means of a standard t-test based on the aggregated data. To study the effects of such a test (denoted by T_x in the following), we compare its asymptotic power func-

tion to the asymptotic power function of the t-test T based on the non-aggregated data.

First note that the test statistic of T is

$$t := \frac{\hat{\beta}}{\sqrt{\hat{\sigma}_\epsilon^2 / S_x^2}} \sqrt{n} \ , \tag{27}$$

where $\hat{\sigma}_\epsilon^2$ is the estimate of σ_ϵ^2 based on the non-aggregated data. It is known that if $\beta \neq 0$, the power function of T converges to 1 as $n \to \infty$. Therefore, in order to compare the asymptotic power functions of T_x and T, we do this for "local alternatives" $\beta = \beta_0/\sqrt{n}$. As $\hat{\sigma}_\epsilon^2 \to \sigma_\epsilon^2$ and $S_x^2 \to \sigma_x^2$, it follows that

$$t \approx \frac{\hat{\beta}}{v} \sqrt{n} = \frac{\hat{\beta} - \frac{\beta_0}{\sqrt{n}}}{v} \sqrt{n} + \frac{\beta_0}{v} \ . \tag{28}$$

Thus, if $\beta_0/\sqrt{n}$ is the true slope parameter,

$$t \to \mathrm{N}\Big(\frac{\beta_0}{v}, 1\Big) \ . \tag{29}$$

Now consider the t-test T_x based on the n/A *distinguishable* data values that are aggregated with respect to X. The test statistic of T_x becomes

$$t_x := \frac{\tilde{\beta}_x}{\sqrt{\tilde{\sigma}_{\epsilon,x}^2 / S_{\tilde{x}_x}^2}} \sqrt{\frac{n}{A}} \ , \tag{30}$$

where $\tilde{\sigma}^2_{\epsilon,x}$ is the naive estimator of σ^2_ϵ based on the aggregated data. As $\tilde{\sigma}^2_{\epsilon,x} \to (1/A)\sigma^2_\epsilon$ and $S^2_{\tilde{x}_x} \to \sigma^2_x$, (see Schmid et al. (2005)), t_x is asymptotically equal to

$$t_x \approx \frac{\tilde{\beta}_x}{\sqrt{\sigma^2_\epsilon/(A\sigma^2_x)}}\sqrt{\frac{n}{A}} = \frac{\tilde{\beta}_x}{v}\sqrt{n} = \frac{\sqrt{n}(\tilde{\beta}_x - \frac{\beta_0}{\sqrt{n}})}{v} + \frac{\beta_0}{v} \ . \tag{31}$$

With the help of (31), we can obtain the asymptotic distribution of t_x:

Theorem 4. *Under the local alternative $\beta = \beta_0/\sqrt{n}$, t_x is asymptotically normally distributed with mean β_0/v and variance one.*

Proof: As $\sqrt{n}(\tilde{\beta}_x - \hat{\beta}) \to 0$ (see Theorem 1), $\sqrt{n}(\tilde{\beta}_x - \frac{\beta_0}{\sqrt{n}})$ is asymptotically normally distributed with mean 0 and variance v^2. Therefore, by (31),

$$t_x \xrightarrow{d} N\Big(\frac{\beta_0}{v}, 1\Big) \ . \tag{32}$$

Comparing (29) to (32), we see that T and T_x asymptotically have the same power functions. Note that this is only true if the n/A distinguishable data values are used for testing. On the other hand, it follows from (31) and (32) that the null hypothesis "$H_0 : \beta = 0$" would be rejected too often if *all* n data values were used for the t-test. The test would not meet the nominal significance level.

4.2 Microaggregation with Respect to Y

Now consider the case where the data have been microaggregated with respect to the dependent variable Y. Similarly to the previous section, we compare the asymptotic power function of the naive t-test T_y based on the n/A distinguishable aggregated data values to the asymptotic power function of T.

First note that the test statistic of T_y is

$$t_y := \frac{\tilde{\beta}_y}{\sqrt{\tilde{\sigma}^2_{\epsilon,y}/S^2_{\tilde{x}_y}}}\sqrt{\frac{n}{A}}\ , \tag{33}$$

where $\tilde{\sigma}^2_{\epsilon,y}$ is the naive estimate of σ^2_ϵ based on the aggregated data. As $\tilde{\sigma}^2_{\epsilon,y} \to (1/A)f(\rho)\sigma^2_\epsilon$ and $S^2_{\tilde{x}_y} \to \sigma^2_x/f(\rho)$, see Schmid et al. (2005), t_y is asymptotically equal to

$$t_y \approx \frac{\tilde{\beta}_y}{\sqrt{\sigma^2_\epsilon f(\rho)^2/(A\sigma^2_x)}}\sqrt{\frac{n}{A}} = \frac{\tilde{\beta}_y}{vf(\rho)}\sqrt{n} = \frac{\tilde{\beta}_y - \frac{\beta_0}{\sqrt{n}}f(\rho)}{vf(\rho)}\sqrt{n} + \frac{\beta_0}{v}\ . \tag{34}$$

To obtain the distribution of $\frac{\tilde{\beta}_y - \frac{\beta_0}{\sqrt{n}}f(\rho)}{vf(\rho)}\sqrt{n}$, we make use of the following lemma:

Lemma 3. *$\big(\tilde{\beta}_y - \beta f(\rho)\big)$ is asymptotically equivalent to $f(\rho)\big(\hat{\beta} - \beta + \beta K\big)$, where*

$$K := \frac{f(\rho)}{\sigma^2_x}\Big(\big(\frac{1}{A}S^2_\delta - S^2_{\tilde{\delta}_y}\big) - \big(1 - \frac{1}{A}\big)\sigma^2_x\big(\hat{\rho}^2 - \rho^2\big) \Big) \tag{35}$$

and $\sqrt{n}K \xrightarrow{d} N(0, \sigma^2_K)$.

Proof: See appendix.

With the help of Lemma 3, we can obtain the asymptotic distribution of t_y:

Theorem 5. *Under the local alternative $\beta = \beta_0/\sqrt{n}$, t_y is asymptotically normally distributed with mean β_0/v and variance one.*

Proof: Denote by σ^2_K the variance of K and by $\sigma_{\hat{\beta}K}$ the covariance of $\hat{\beta}$ and K. Assuming $\beta_0/\sqrt{n}$ to be the true slope parameter, (34) together with Lemma 3 yields

$$
\begin{aligned}
t_y &\approx \frac{\sqrt{n}\big(\tilde{\beta}_y - \frac{\beta_0}{\sqrt{n}} f(\rho)\big)}{v f(\rho)} + \frac{\beta_0}{v} \\
&\approx \frac{\sqrt{n}\big(\hat{\beta} - \frac{\beta_0}{\sqrt{n}} + \frac{\beta_0}{\sqrt{n}} K\big)}{v} + \frac{\beta_0}{v} \\
&\xrightarrow{d} \mathrm{N}\Big(\frac{\beta_0}{v}, \big(1 + \frac{\beta_0^2}{v^2 n}\sigma_K^2 + 2\frac{\beta_0}{v\sqrt{n}}\sigma_{\hat{\beta}K}\big)\Big) \\
&\xrightarrow{d} \mathrm{N}\Big(\frac{\beta_0}{v}, 1\Big) \ . \qquad (36)
\end{aligned}
$$

Comparing (36) to (29), we see that T_y and T have the same power functions asymptotically. Therefore, just as in the case where the data are aggregated with respect to X, we obtain an unbiased t-test by applying the standard t-test to the n/A distinguishable data values. Again it may be noted that if all n aggregated data values were used, the resulting standard t-test would not meet the nominal significance level.

Note that in order to test the hypothesis "$H_0 : \beta = 0$", we do not need to correct for the bias of $\tilde{\beta}_y$. As shown above, it is sufficient to compute $\tilde{\beta}$, $\tilde{\sigma}^2_{\epsilon,y}$ and $S^2_{\tilde{x}_y}$ to obtain an (asymptotically) unbiased t-test.

5 Simulations

5.1 Finite Sample Variances of $\tilde{\beta}_x$, $\tilde{\beta}_y$, and $\tilde{\beta}_{y,c}$

In this section, we check to which extent the asymptotic results of section 3 hold in realistic data situations. For this purpose, we computed the variances of $\tilde{\beta}_x$, $\tilde{\beta}_y$, and $\tilde{\beta}_{y,c}$ for various n and various values of β. The residual standard deviation σ_ϵ was

set to three, A was set to three as well. α was set to one and $X \sim \mathrm{N}(0, 2^2)$. Fig. 1 shows the variance of $\sqrt{n}\tilde{\beta}_x$, based on 1000 replications. In addition, Fig. 1 includes the mean of the estimated asymptotic variance based on (9). We see that if n is small, the estimated asymptotic variance of $\sqrt{n}\tilde{\beta}_x$ is smaller than its true value. As n increases, the approximation of the variance of $\tilde{\beta}_x$ works as it should: $\mathrm{var}(\sqrt{n}\tilde{\beta}_x)$ is almost identical to the true variance and to the variance of $\sqrt{n}\hat{\beta}$. Furthermore, as expected, $\mathrm{var}(\sqrt{n}\tilde{\beta}_x)$ does not depend on β.

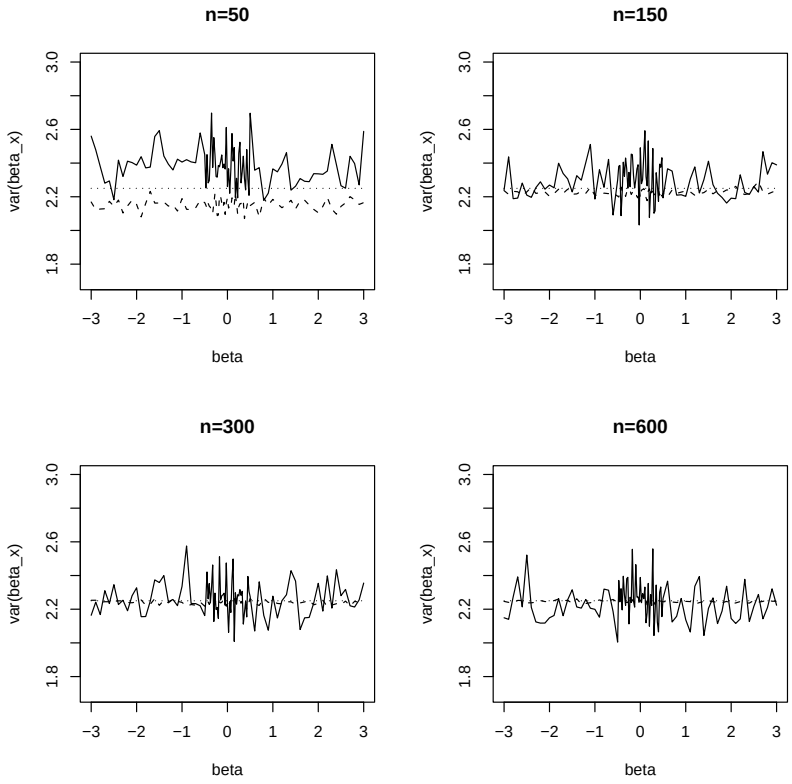

Figure 1: Variance curves of $\sqrt{n}\tilde{\beta}_x$ (solid line = true variance, dashed line = estimated asymptotic variance, dotted line = v^2)

In the same way, we computed the variance of $\sqrt{n}\tilde{\beta}_y$ for various n. Fig. 2 shows the variance of $\sqrt{n}\tilde{\beta}_y$, based on 1000 replications. In addition, Fig. 2 includes the mean of the estimated asymptotic variance based on (21). We see that if n is small, the estimated asymptotic variance of $\sqrt{n}\tilde{\beta}_y$ differs from its true value. As n increases, the approximation of the variance of $\sqrt{n}\tilde{\beta}_y$ works as it should: $\mathrm{var}(\sqrt{n}\tilde{\beta}_y)$ is almost identical to the true variance. We also see that, contrary to microaggregation with respect to X, $\mathrm{var}(\sqrt{n}\tilde{\beta}_y)$, does depend on β. It has its extreme value at $\beta = 0$ and flattens as $|\beta| \to \infty$.

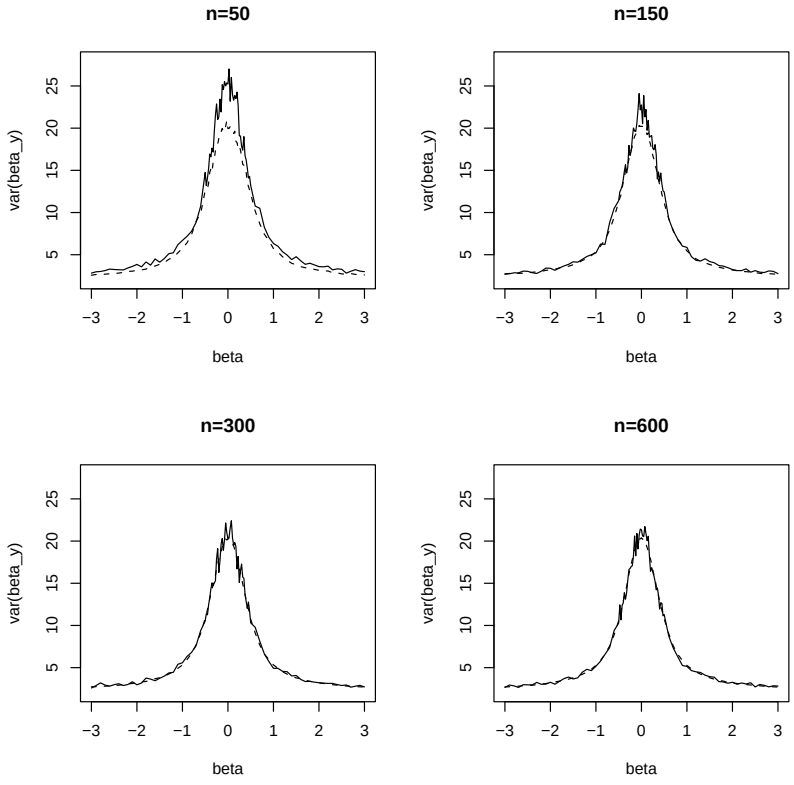

Figure 2: Variance curves of $\sqrt{n}\tilde{\beta}_y$ (solid line = true variance, dashed line = estimated asymptotic variance)

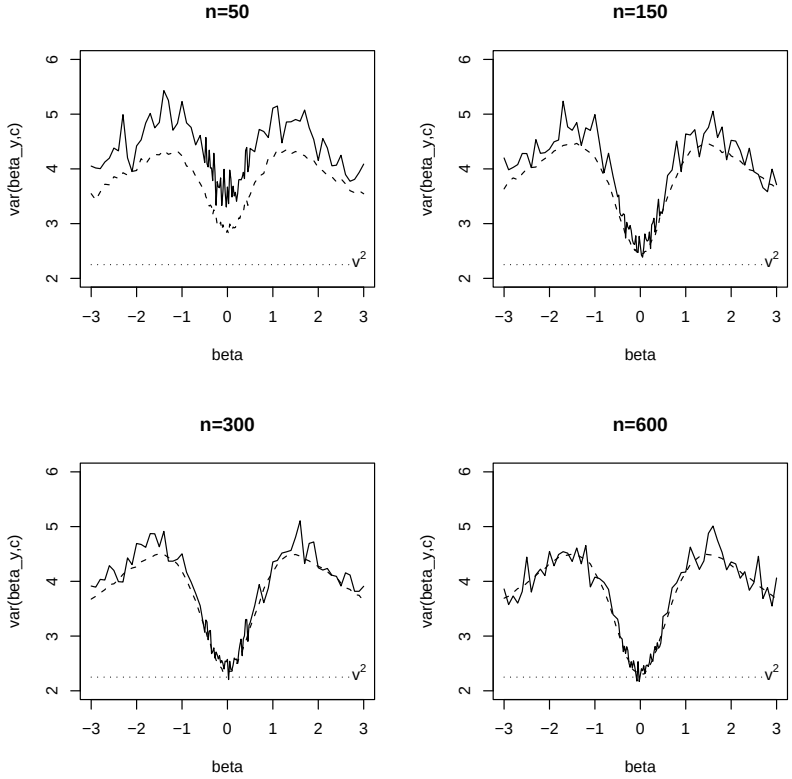

Figure 3: Variance curves of $\sqrt{n}\tilde{\beta}_{y,c}$ (solid line = true variance, dashed line = estimated asymptotic variance)

Fig. 3 shows the variance of the corrected least squares estimator $\sqrt{n}\tilde{\beta}_{y,c}$, together with the mean of the estimated asymptotic variance based on (25). Obviously, if n is small, the asymptotic variance of $\sqrt{n}\tilde{\beta}_{y,c}$ is smaller than its true variance. As n increases, the approximation of the variance of $\sqrt{n}\tilde{\beta}_{y,c}$ works as it should: The mean of $\text{var}(\sqrt{n}\tilde{\beta}_{y,c})$ is almost identical to the true variance. We also see that, contrary to the variance of $\tilde{\beta}_y$, $\text{var}(\sqrt{n}\tilde{\beta}_{y,c})$ is smallest and equal to v^2, the variance of $\sqrt{n}\hat{\beta}$, when $\beta = 0$. As $|\beta| \to \infty$, $\text{var}(\sqrt{n}\tilde{\beta}_{y,c})$ flattens again.

5.2 Confidence Intervals for β

In this section, we study the behavior of the asymptotic confidence intervals (10) and (26) for various n and various values of β. To achieve this, we carried out a simulation study based on 1000 replications. The residual standard deviation σ_ϵ was set to three, A was set to three as well.

First, we performed microaggregation with respect to X. For each replication, a 95% confidence interval for β based on the non-aggregated data was computed. Moreover, we computed the corresponding asymptotic 95% confidence interval based on (10). The results are shown in Table 1. The third column of Table 1 shows the mean width of the 1000 confidence intervals based on the original data. The mean width of the 1000 asymptotic confidence intervals based on (10) is presented in column five of Table 1. We see that, as suggested by (10), the width of the asymptotic confidence intervals does not depend on β. Moreover, if n is small, the asymptotic confidence intervals are smaller than the confidence intervals based on the non-aggregated data, which is somewhat surprising. As Fig. 1 suggests, this effect is due to the underestimation of $\text{var}(\tilde{\beta}_x)$ for small n. If n is large, the confidence intervals based on the original data and the confidence intervals based on the aggregated data have almost equal length. Columns four and six in Table 1 show the coverage rates of the confidence intervals based on 1000 replications. Apparently, for any n, the coverage rates of the confidence intervals based on the aggregated data are lower than the coverage rates of the asymptotic confidence intervals based on the non-aggregated data. If n is small, the coverage rates of the asymptotic intervals are considerably smaller than the degree of confidence (which is 95%). For large n the

		Original data		Aggregated data	
n	β	Width of CI	Cov. rate	Width of CI	Cov. rate
50	0	0.845	0.940	0.800	0.904
	1	0.836	0.952	0.793	0.936
	2	0.846	0.965	0.780	0.939
	5	0.846	0.934	0.801	0.915
150	0	0.484	0.961	0.475	0.955
	1	0.484	0.939	0.476	0.934
	2	0.486	0.936	0.478	0.917
	5	0.483	0.964	0.474	0.958
300	0	0.340	0.951	0.336	0.951
	1	0.340	0.951	0.336	0.949
	2	0.341	0.936	0.338	0.930
	5	0.341	0.946	0.339	0.942
600	0	0.240	0.949	0.240	0.943
	1	0.240	0.955	0.239	0.951
	2	0.241	0.947	0.239	0.945
	5	0.240	0.941	0.239	0.939
1200	0	0.170	0.952	0.170	0.952
	1	0.170	0.954	0.169	0.949
	2	0.170	0.956	0.169	0.952
	5	0.170	0.940	0.170	0.937

Table 1: Confidence intervals for β (Microaggregation with respect to X)

difference is negligible.

Next, we performed microaggregation with respect to Y. For each replication, a 95% confidence interval for β based on the non-aggregated data was computed. Moreover, we computed an asymptotic 95% confidence interval based on (5) and (26). The results are shown in Table 2. The third column of Table 2 shows the mean width of the 1000 confidence intervals based on the original data. The mean width of the 1000 asymptotic confidence intervals based on (26) is presented in column five of Table 2.

		Original data		Aggregated data	
n	β	Width of CI	Cov. rate	Width of CI	Cov. rate
50	0	0.836	0.943	0.909	0.952
	1	0.834	0.933	1.122	0.904
	2	0.839	0.939	1.010	0.906
	5	0.839	0.935	0.894	0.885
150	0	0.483	0.948	0.497	0.945
	1	0.484	0.945	0.655	0.945
	2	0.483	0.959	0.656	0.930
	5	0.483	0.948	0.541	0.939
300	0	0.340	0.951	0.345	0.954
	1	0.340	0.952	0.462	0.951
	2	0.340	0.950	0.467	0.946
	5	0.342	0.956	0.387	0.937
600	0	0.240	0.942	0.242	0.943
	1	0.240	0.949	0.327	0.948
	2	0.240	0.951	0.331	0.939
	5	0.240	0.943	0.273	0.941
1200	0	0.170	0.949	0.171	0.948
	1	0.170	0.949	0.231	0.952
	2	0.170	0.947	0.235	0.944
	5	0.170	0.949	0.193	0.955

Table 2: Confidence intervals for β (Microaggregation with respect to Y)

We see that, contrary to microaggregation with respect to X, the width of the asymptotic confidence intervals depends on β. As suggested by Fig. 3, the asymptotic confidence intervals are smallest when $\beta = 0$. As β increases, they become larger. For very large values of β (here, $\beta = 5$) the asymptotic confidence intervals become smaller again. Table 2 also shows that, for any n, the confidence intervals based on the non-aggregated data are smaller than the asymptotic confidence intervals based on the aggregated data. However, if $\beta = 0$ and n is large, this difference almost disappears. Concerning the coverage rates, we see that if n is small, the asymptotic intervals do not keep the degree of confidence (which is 95%) except for $\beta = 0$.

5.3 T-Tests

In this section, we check whether the results of section 4 hold in realistic data situations. First, we performed microaggregation with respect to X. To estimate the power function of T_x, we carried out a simulation study based on 500 replications. As before, we set $\alpha = 1$, $\sigma_\epsilon = 3$, and $A = 3$. For each replication, we carried out a t-test based on the n/A distinguishable aggregated data. Moreover, we carried out t-tests based on all n aggregated data and on the non-aggregated data. Next, in order to estimate the power functions of these tests, we computed the proportion of tests that rejected the null hypothesis "$H_0 : \beta = 0$". The significance level was chosen to be $\alpha = 0.05$.

Fig. 4 shows the estimated power functions for four values of n ($n = 50$, $n = 150$, $n = 300$, and $n = 600$). We see that even for small sample sizes, the power function of T_x is a very good approximation of the power function of T. As expected, the test based on all n aggregated data does not meet the nominal significance level: H_0 is rejected in more than 5% of all cases.

Next, we performed microaggregation with respect to Y. Again, for each replication, we carried out a t-test based on the n/A distinguishable aggregated data. Moreover, we carried out t-tests based on all n aggregated data and on the non-aggregated data. As before, in order to estimate the power functions of these tests, we computed the proportion of tests that rejected the null hypothesis "$H_0 : \beta = 0$". The significance level was chosen to be $\alpha = 0.05$.

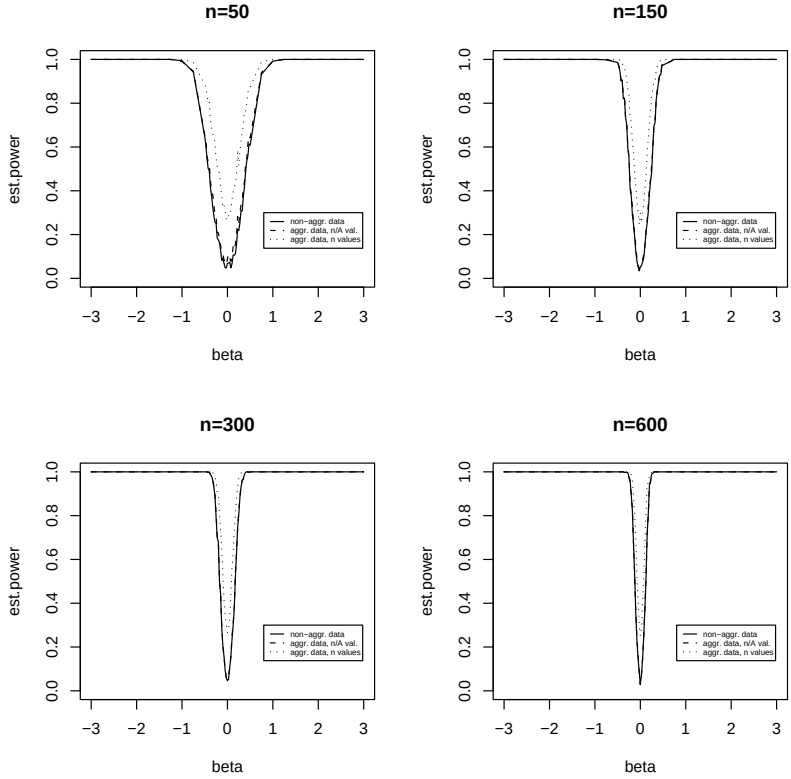

Figure 4: Power of the standard t-test (data aggregated with respect to X)

Fig. 5 shows the estimated power functions for four values of n ($n = 50$, $n = 150$, $n = 300$, and $n = 600$). The results are basically the same as when microaggregation with respect to X is performed: Even for small sample sizes, the power function of T_x is a very good approximation of the power function of T. Again, the test based on all n aggregated data does not meet the nominal significance level: H_0 is rejected in more than 5% of all cases.

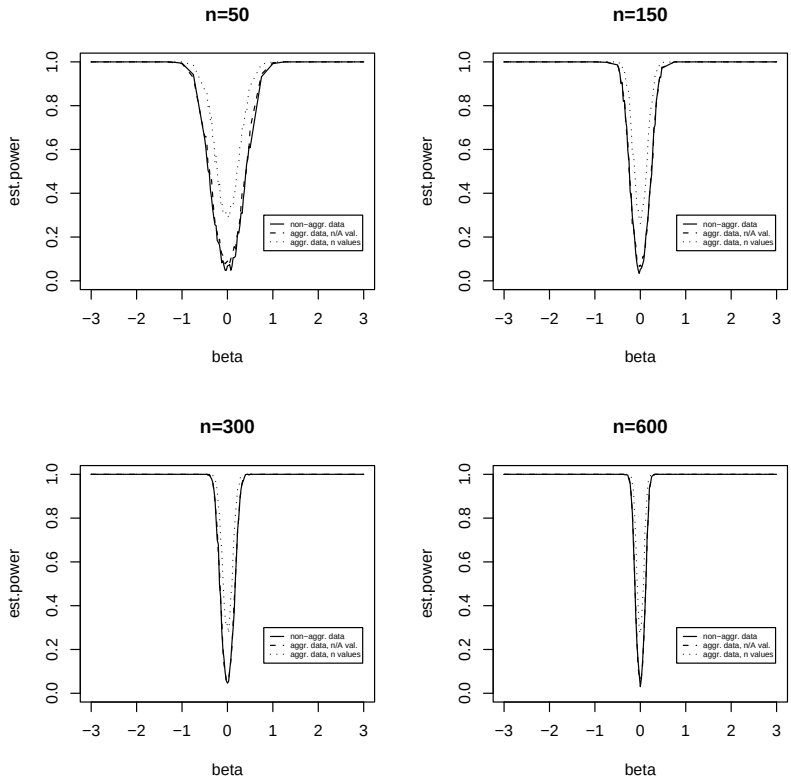

Figure 5: Power of the standard t-test (data aggregated with respect to Y)

6 Conclusion

Microaggregation clearly has an effect on both the disclosure risk of a data file and its analytical validity. Over the last years, scientific research has mainly been concerned with the former issue. In contrast, there has not been much work to date that describes the analytic properties of a masked data set. However, it is vitally important to know in which way statistical analysis is affected by anonymization techniques. In this paper, we focused on the effects of microaggregation on the estimation of a linear model, one of the most frequently used statistical methods.

The main results are:

1. Concerning microaggregation with respect to X, it is possible to derive an asymptotic variance formula for the naive estimate $\tilde{\beta}_x$. This formula is equal to the asymptotic variance of the least squares estimate $\hat{\beta}$ based on the non-aggregated data. In particular, it is independent of β.

2. Concerning microaggregation with respect to Y, asymptotic variance formulas for the naive least squares estimate $\tilde{\beta}_y$ and for the corrected least squares estimate $\tilde{\beta}_{y,c}$ can be derived by means of the delta method.

3. With the help of the asymptotic variance formulas $\mathrm{var}(\tilde{\beta}_x)$ and $\mathrm{var}(\tilde{\beta}_{y,c})$, asymptotic confidence intervals for β can be constructed. If the data are aggregated with respect to X, the width of these intervals does not depend on β.

4. The simulation study in section 5 shows that for $n \geq 150$, the asymptotic

variance formulas are a good approximation of the true variances of $\tilde{\beta}_x$, $\tilde{\beta}_y$, and $\tilde{\beta}_{y,c}$. The simulations also indicate that $\text{var}(\tilde{\beta}_y)$ and $\text{var}(\tilde{\beta}_{y,c})$ depend on the value of β.

5. The asymptotic confidence intervals derived in section 3 show satisfactory coverage rates for $n \geq 150$. Moreover, if the data are microaggregated with respect to X, the width of the asymptotic intervals is almost equal to the width of the intervals based on the non-aggregated data, at least for large n. If the data are aggregated with respect to Y, the width of the asymptotic confidence intervals depends on β. In addition, for $\beta \neq 0$, the intervals are larger than the intervals based on the non-aggregated data.

6. The power function of the t-test for the slope parameter β based on the non-aggregated data can asymptotically be preserved by carrying out a naive t-test based on the n/A distinguishable aggregated data values. This result holds for both microaggregation with respect to X and microaggregation with respect to Y. In contrast, the naive t-test with all n aggregated data values does not meet the nominal significance level.

7. The simulation study in section 5 shows that even for small n, the asymptotic power functions derived in section 4 are very good approximations of the power function of the t-test based on the non-aggregated data.

Together with the results presented in Schmid et al. (2005), we have developed an asymptotic theory for estimating a simple linear model based on microaggregated data. Of course, this theory relies on the assumption that a leading variable is used

for grouping the data.

Future research topics include

- multiple regression: We have developed our theory for a simple linear model with one covariate. Clearly, this model can be extended to a set of more than one covariate. It is therefore necessary to investigate the effects of microaggregation on multiple linear regression.
- the inclusion of discrete covariates: Microaggregation is primarily used for masking continuous variables. Nevertheless, if a linear model includes (non-anonymized) discrete covariates, it is highly likely that the parameter estimates of these discrete covariates are affected by microaggregation of the continuous covariates. Therefore, methods for quantifying a possible bias of the least squares estimates have to be developed.
- a sensitivity analysis: The theory we have developed is based on the assumption that the covariate X is normally distributed (at least if the data are microaggregated with respect to Y). In practice, however, this assumption might not always be justified. It is therefore necessary to analyze the sensitivity of the bias and variance formulas to deviations from normality.

Acknowledgements

We gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (German Science Foundation), and we thank Daniel Rost for helpful discussions.

Appendix - Proofs

Lemma 1.

a) Denote by S_y^2 *the empirical variance of* y*. Then* $\sqrt{n}(S_{\tilde{y}_y}^2 - S_y^2)$ *converges in probability to 0.*

b) Denote by $S_{y\delta}$ *the empirical covariance of* Y *and* δ *in model (6). Analogously, denote by* $S_{\tilde{y}_y\tilde{\delta}_y}$ *the empirical covariance of* $\tilde{y}_y$ *and* $\tilde{\delta}_y$ *Then,* $\sqrt{n}(S_{\tilde{y}_y\tilde{\delta}_y} - S_{y\delta})$ *converges in probability to 0.*

c) Denote by S_{xy} *the empirical covariance of* x *and* y*. Then* $\sqrt{n}(S_{\tilde{x}_y\tilde{y}_y} - S_{xy})$ *converges in probability to 0.*

d) Denote by S_x^2 *the empirical variance of* x*. Then* $S_{\tilde{x}_y}^2 - S_x^2$ *is asymptotically equivalent to* $S_{\tilde{\delta}_y}^2 - S_\delta^2$*.*

e) For $n \to \infty$*,* $\sqrt{n}(S_{\tilde{\delta}_y}^2 - \frac{1}{A}S_\delta^2)$ *converges to a normal distribution with zero mean and variance* $2\frac{A-1}{A^2}\sigma_\delta^4 = 2\frac{A-1}{A^2}\sigma_x^4(1-\rho^2)^2$*.*

f) Consider the equation

$$X_i = \hat{\beta}^* Y_i + \hat{\delta}_i \ , \ i = 1, \ldots, n \ , \tag{37}$$

where $\hat{\beta}^*$ *is the least squares estimate based on the non-aggregated data and* $\hat{\delta}_i := X_i - \hat{\beta}^* Y_i$ *is the corresponding residual term. Then,* $S_{\hat{\delta}}^2$ *is asymptotically equivalent to the empirical variance* $S_{\hat{\delta}}^2$ *of* $\hat{\delta}_1, \ldots, \hat{\delta}_n$*.*

Proof of a): We prove a) for $A = 2$. For $A > 2$, the proof is analogous. Without loss of generality, we set $\mu_y = 0$ and $\sigma_y^2 = 1$. Denote by $S_{y,W}^2$ and $S_{y,B}^2$ the within-

groups variance and the between-groups variance of Y respectively. By definition, $S^2_{y,B} = S^2_{\tilde{y}_y}$. Moreover, denote by $Y_{i:n}$ the i-th order statistic of the sample variables $Y_1, \ldots, Y_n$.

Now, as

$$\sqrt{n}\,(S^2_y - S^2_{\tilde{y}_y}) \;=\; \sqrt{n}\,S^2_{y,W} \;\leq\; \frac{1}{\sqrt{n}}\,\frac{1}{2}\sum_{i=2}^{n}(Y_{i:n} - Y_{(i-1):n})^2 \;, \tag{38}$$

it is sufficient to show that $1/\sqrt{n}\;\sum_{i=2}^{n}(Y_{i:n} - Y_{(i-1):n})^2$ converges in probability to 0. Define $b_n := n^{1/8}$ and $M_n := \max\{|Y_1|, \ldots, |Y_n|\}$.

Next, for any $\epsilon > 0$, we consider the events

$$A \;:=\; \left\{ \frac{1}{\sqrt{n}}\sum_{i=2}^{n}(Y_{i:n} - Y_{(i-1):n})^2 > \epsilon \right\} \;, \tag{39}$$

$$B \;:=\; \left\{ M_n \leq b_n \right\} \;. \tag{40}$$

We have to show that $\mathrm{P}(A) \to 0$. Clearly, the following inequality holds:

$$\begin{aligned} \mathrm{P}(A) \;&=\; \mathrm{P}(A \cap B) + \mathrm{P}(A \cap \bar{B}) \\ &\leq\; \mathrm{P}(A \cap B) + \mathrm{P}(\bar{B}) \;. \end{aligned} \tag{41}$$

Now, by showing that the probabilities $\mathrm{P}(A \cap B)$ and $\mathrm{P}(\bar{B})$ both converge to 0, we can prove part a) of Lemma 1.

Let us first consider the event $A \cap B$ in equation (41): Under this event, we have

$$\begin{aligned}
\epsilon &< \frac{1}{\sqrt{n}} \sum_{i=2}^{n} (Y_{i:n} - Y_{(i-1):n})^2 \\
&= \frac{1}{\sqrt{n}} \sum_{i:|Y_{i:n} - Y_{(i-1):n}| \le 1} (Y_{i:n} - Y_{(i-1):n})^2 \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i:|Y_{i:n} - Y_{(i-1):n}| > 1} (Y_{i:n} - Y_{(i-1):n})^2 \\
&\le \frac{1}{\sqrt{n}} \Big(\max_i \{Y_i\} - \min_i \{Y_i\} \Big) + 2b_n \cdot \frac{1}{\sqrt{n}} (2b_n)^2 \\
&\le \frac{1}{\sqrt{n}} 2b_n + \frac{1}{\sqrt{n}} 8b_n^3 \ . \qquad (42)
\end{aligned}$$

It follows that

$$\mathrm{P}(A \cap B) \ \le \ \mathrm{P}\Big(\frac{1}{\sqrt{n}} 2b_n + \frac{1}{\sqrt{n}} 8b_n^3 > \epsilon \Big) \ = \ \mathrm{P}\big(2n^{-3/8} + 8n^{-1/8} > \epsilon \big) \ . \qquad (43)$$

Therefore, $\mathrm{P}(A \cap B) \to 0$.

Next, we consider the event $\bar{B}$ in equation (41). As Y is normally distributed, we have

$$\begin{aligned}
\mathrm{P}(\bar{B}) &= \mathrm{P}\big(\min_i \{Y_i\} < -b_n \cup \max_i \{Y_i\} > b_n \big) \\
&\le \mathrm{P}\big(\min_i \{Y_i\} < -b_n \big) + \mathrm{P}\big(\max_i \{Y_i\} > b_n \big) \\
&= 2\left(1 - \Phi(b_n)^n\right) \ , \qquad (44)
\end{aligned}$$

where $\Phi(\cdot)$ denotes the normal cumulative probability function. Now, if n is large, $\Phi(b_n)$ can be approximated by

$$\begin{aligned}
\Phi(b_n) &\approx 1 - (2\pi)^{-1/2} \frac{1}{b_n} \exp(-b_n^2/2) \\
&= 1 - (2\pi)^{-1/2} n^{-1/8} \exp(-n^{1/4}/2) \ =: \ R_n \ , \qquad (45)
\end{aligned}$$

in the sense that $\frac{1-\Phi(b_n)}{1-R_n} \to 1$ (see Johnson et al. (1994)). Therefore, as

$$R_n^n \;=\; \Big(1 - \frac{(2\pi)^{-1/2} n^{7/8} \exp(-n^{1/4}/2)}{n}\Big)^n \;\to\; \exp(0) \;=\; 1 \;, \tag{46}$$

the right side of (44) converges to 0.

Proof of b): We prove b) for $A = 2$. For $A > 2$, the proof is analogous. Without loss of generality, we set $\mu_y = 0$ and $\sigma_y^2 = 1$. Denote by $\delta_{[i]}$ the error variable associated with $Y_{i:n}$.

First of all, as

$$S_{y\delta} - S_{\tilde{y}_y \tilde{\delta}_y} \;=\; \frac{1}{2n}\Big((Y_{2:n} - Y_{1:n})(\delta_{[2]} - \delta_{[1]}) + \ldots + (Y_{n:n} - Y_{(n-1):n})(\delta_{[n]} - \delta_{[n-1]})\Big) \;, \tag{47}$$

it is sufficient to show that $1/\sqrt{n} \sum_{i=2,4,\ldots,n} (Y_{i:n} - Y_{(i-1):n}) \, |\delta_{[i]} - \delta_{[i-1]}| \to 0$. Define $u_i := \delta_{[i]} - \delta_{[i-1]}$, $i = 2, 4, \ldots, n$. It follows that the u_i are independent normally distributed random variables having mean 0. Without loss of generality, we assume them to be *standard* normal.

Moreover, we define $b_n = c_n := n^{1/8}$, $M_{y,n} := \max\{|Y_1|, \ldots, |Y_n|\}$, and $M_{u,n} := \max\{|u_2|, \ldots, |u_n|\}$.

Next, for any $\epsilon > 0$, consider the events

$$\begin{aligned} A &:= \{\frac{1}{\sqrt{n}} \sum_{i=2,4,\ldots,n} (Y_{i:n} - Y_{(i-1):n})\,|u_i| > \epsilon\}\ , \\ B &:= \{M_{y,n} \leq b_n\}\ , \\ C &:= \{M_{u,n} \leq c_n\}\ . \end{aligned} \tag{48}$$

We have to show that $\mathrm{P}(A) \to 0$. Clearly, the following inequality holds:

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(A \cap B \cap C) + \mathrm{P}(A \cap B \cap \bar{C}) \\ &\quad + \mathrm{P}(A \cap \bar{B} \cap \bar{C}) + \mathrm{P}(A \cap \bar{B} \cap C) \\ &\leq \mathrm{P}(A \cap B \cap C) + \mathrm{P}(\bar{C}) \\ &\quad + \mathrm{P}(\bar{C}) + \mathrm{P}(\bar{B})\ . \end{aligned} \tag{49}$$

Now, by showing that each of the probabilities $\mathrm{P}(A \cap B \cap C)$, $\mathrm{P}(\bar{B})$, and $\mathrm{P}(\bar{C})$ converges to 0, we are able to prove part b) of Lemma 1.

Let us first consider the event $A \cap B \cap C$. Under this event, we have

$$\begin{aligned} \epsilon &< \frac{1}{\sqrt{n}} \sum_{i=2,4,\ldots,n} (Y_{i:n} - Y_{(i-1):n})\,|u_i| \\ &\leq \frac{1}{\sqrt{n}} c_n \Big(\max_i\{Y_i\} - \min_i\{Y_i\} \Big) \\ &\leq \frac{1}{\sqrt{n}}\, 2\, b_n c_n\ . \end{aligned} \tag{50}$$

It follows that

$$\mathrm{P}(A \cap B \cap C) \;\leq\; \mathrm{P}\Big(\frac{1}{\sqrt{n}} 2 b_n c_n > \epsilon\Big) \;=\; \mathrm{P}(2n^{-1/4} > \epsilon)\;. \tag{51}$$

Therefore $\mathrm{P}(A \cap B \cap C) \to 0$.

Next, we consider $\mathrm{P}(\bar{B})$ and $\mathrm{P}(\bar{C})$. From (44) we obtain

$$\begin{aligned} \mathrm{P}(\bar{B}) &\leq 2\big(1 - \Phi(b_n)^{n/2}\big)\;, & (52)\\ \mathrm{P}(\bar{C}) &\leq 2\big(1 - \Phi(c_n)^{n/2}\big) \;=\; 2\,\big(1 - \Phi(b_n)^{n/2}\big). & (53) \end{aligned}$$

As (45) can again be used to approximate $\Phi(b_n)$, it follows that both $\mathrm{P}(\bar{B})$ and $\mathrm{P}(\bar{C})$ converge to 0.

Proof of c): As

$$\begin{aligned} S_{xy} &= \beta^{*2} S_y^2 + S_{y\delta}\;, & (54)\\ S_{\tilde{x}_y \tilde{y}_y} &= \beta^{*2} S_{\tilde{y}_y}^2 + S_{\tilde{y}_y \tilde{\delta}_y}\;, & (55) \end{aligned}$$

we obtain

$$\sqrt{n}\,(S_{\tilde{x}_y \tilde{y}_y} - S_{xy}) \;=\; \sqrt{n}\,\Big(\beta^{*2}(S_{\tilde{y}_y}^2 - S_y^2) + S_{\tilde{y}_y \tilde{\delta}_y} - S_{y\delta}\Big)\;. \tag{56}$$

Because of a) and b), (56) converges in probability to 0.

Proof of d): As

$$S_x^2 = \beta^{*2} S_y^2 + 2\beta^* S_{y\delta} + S_\delta^2 \ , \tag{57}$$

$$S_{\tilde{x}_y} = \beta^{*2} S_{\tilde{y}_y}^2 + 2\beta^* S_{\tilde{y}_y \tilde{\delta}_y} + S_{\tilde{\delta}_y}^2 \ , \tag{58}$$

we obtain

$$\sqrt{n}\,(S_x^2 - S_{\tilde{x}_y}^2) = \sqrt{n}\left(\beta^{*2}(S_y^2 - S_{\tilde{y}_y}^2) + 2\beta^*(S_{y\delta} - S_{\tilde{y}_y\tilde{\delta}_y}) + S_\delta^2 - S_{\tilde{\delta}_y}^2\right) . \tag{59}$$

Hence d) follows from a) and b).

Proof of e): First of all, $S_{\tilde{\delta}_y}^2 - \frac{1}{A} S_\delta^2$ can be written as

$$\begin{aligned} S_{\tilde{\delta}_y}^2 - \frac{1}{A} S_\delta^2 &= \frac{A}{n}\Big(\big(\frac{1}{A}\sum_{i=1}^{A} \delta_{[i]}\big)^2 + \big(\frac{1}{A}\sum_{i=A+1}^{2A} \delta_{[i]}\big)^2 + \dots \Big) \\ &\quad - \frac{1}{nA}\left(\delta_{[1]}^2 + \dots + \delta_{[n]}^2\right) \\ &= \frac{2}{A^2}\,\frac{A}{n}\sum_{k=1}^{n/A} S_k \ , \end{aligned} \tag{60}$$

where

$$S_k := \sum_{\substack{i<j \\ i,j\in\{(k-1)A+1,\dots,kA\}}} \delta_{[i]}\delta_{[j]} \ . \tag{61}$$

Define $\bar{S}_s := \frac{A}{n}\sum_{k=1}^{n/A} S_k$. Now, as $\delta_{[1]}, \dots, \delta_{[n]}$ are independent identically distributed with zero mean and variance σ_δ^2, and therefore $S_1, \dots, S_{n/A}$ are also iid with zero

mean and variance $\frac{A(A-1)}{2}\sigma_\delta^4$, we have

$$\sqrt{n}\Big(S^2_{\tilde{\delta}_y} - \frac{1}{A}S^2_\delta\Big) = \frac{2}{A^{3/2}}\sqrt{\frac{n}{A}}\bar{S}_s \rightarrow N\Big(0, \frac{4}{A^3}\frac{A(A-1)}{2}\sigma_\delta^4\Big) = N\Big(0, 2\,\frac{A-1}{A^2}\sigma_\delta^4\Big)\ . \tag{62}$$

Proof of f): As $\hat{\delta}_i - \delta_i = (\beta^* - \hat{\beta}^*)Y_i$, it follows that $\delta_i = \hat{\delta}_i - (\beta^* - \hat{\beta}^*)Y_i$. Therefore,

$$S^2_\delta = S^2_{\hat{\delta}} + (\beta^* - \hat{\beta}^*)^2 S^2_y\ . \tag{63}$$

Clearly, $\sqrt{n}(\beta^* - \hat{\beta}^*)^2 S^2_y$ converges to 0.

Therefore,

$$\sqrt{n}(S^2_\delta - S^2_{\hat{\delta}}) \rightarrow 0\ . \tag{64}$$

Note in addition that $S^2_{\hat{\delta}}$ is equal to $S^2_x(1-\hat{\rho}^2)$.

Lemma 2. *$S^2_{\tilde{x}_y}$ is asymptotically equivalent to $\frac{S^2_x}{f(\rho)} + (1-\frac{1}{A})\sigma^2_x(\hat{\rho}^2 - \rho^2) + (S^2_{\tilde{\delta}_y} - \frac{1}{A}S^2_\delta)$.*

Proof: By using parts d) and f) of Lemma 1, we obtain

$$\begin{aligned}
\frac{S_x^2}{f(\rho)} - S_{\tilde{x}_y}^2 &= S_x^2\Big(\frac{1}{A} + \big(1-\frac{1}{A}\big)\rho^2\Big) - S_{\tilde{x}_y}^2 \\
&= S_x^2 - S_{\tilde{x}_y}^2 - S_x^2\big(1-\frac{1}{A}\big)\big(1-\rho^2\big) \\
&\sim S_\delta^2 - S_{\tilde{\delta}_y}^2 - S_x^2\big(1-\frac{1}{A}\big)\big(1-\rho^2\big) \\
&= \frac{1}{A}S_\delta^2 - S_{\tilde{\delta}_y}^2 - \big(1-\frac{1}{A}\big)\big(S_x^2(1-\rho^2) - S_\delta^2\big) \\
&\sim \frac{1}{A}S_\delta^2 - S_{\tilde{\delta}_y}^2 - \big(1-\frac{1}{A}\big)\big(S_x^2(1-\rho^2) - S_{\hat{\delta}}^2\big) \\
&= \frac{1}{A}S_\delta^2 - S_{\tilde{\delta}_y}^2 - \big(1-\frac{1}{A}\big)S_x^2\big(\hat{\rho}^2 - \rho^2\big). \qquad (65)
\end{aligned}$$

Lemma 3. *$(\tilde{\beta}_y - \beta f(\rho))$ is asymptotically equivalent to $f(\rho)(\hat{\beta} - \beta + \beta K)$, where*

$$K := \frac{f(\rho)}{\sigma_x^2}\Big(\big(\frac{1}{A}S_\delta^2 - S_{\tilde{\delta}_y}^2\big) - \big(1-\frac{1}{A}\big)\sigma_x^2\big(\hat{\rho}^2 - \rho^2\big)\Big) \qquad (66)$$

and $\sqrt{n}K \xrightarrow{d} N(0, \sigma_K^2)$.

Proof: First of all, $\sqrt{n}(\tilde{\beta} - \beta f(\rho))$ can be written in the following way:

$$\begin{aligned}
\sqrt{n}\,(\tilde{\beta} - \beta f(\rho)) &= \sqrt{n}\,\Big(\frac{S_{\tilde{x}_y\tilde{y}_y}}{S_{\tilde{x}_y}^2} - \beta f(\rho)\Big) \\
&= \sqrt{n}\,\frac{S_{\tilde{x}_y\tilde{y}_y} - S_{xy}}{S_{\tilde{x}_y}^2} + \sqrt{n}\,\Big(\frac{S_{xy}}{S_x^2}\,\frac{S_x^2}{S_{\tilde{x}_y}^2} - \beta f(\rho)\Big) \\
&= \sqrt{n}\,\frac{S_{\tilde{x}_y\tilde{y}_y} - S_{xy}}{S_{\tilde{x}_y}^2} \\
&\quad + \sqrt{n}\,\Big(\hat{\beta} - \beta + \hat{\beta}\big(\frac{S_x^2}{f(\rho)} - S_{\tilde{x}_y}^2\big)\frac{1}{S_{\tilde{x}_y}^2}\Big) f(\rho) \\
&\approx \sqrt{n}\,\Big(\hat{\beta} - \beta + \hat{\beta}\big(\frac{S_x^2}{f(\rho)} - S_{\tilde{x}_y}^2\big)\frac{1}{S_{\tilde{x}_y}^2}\Big) f(\rho) \qquad (67)
\end{aligned}$$

by part c) of Lemma 1. Since the distribution of $\sqrt{n}(\hat{\beta}-\beta)f(\rho)$ is known from linear model theory, we only consider $\sqrt{n}(\frac{S_x^2}{f(\rho)} - S_{\tilde{x}_y}^2)$. Using Lemma 2, we obtain

$$\begin{aligned}\sqrt{n}\Big(\frac{S_x^2}{f(\rho)} - S_{\tilde{x}_y}^2\Big) &\approx \sqrt{n}\Big(\frac{1}{A}S_\delta^2 - S_{\tilde{\delta}_y}^2 - \big(1-\frac{1}{A}\big)S_x^2(\hat{\rho}^2-\rho^2)\Big) \\ &= \sqrt{n}\,\frac{\sigma_x^2}{f(\rho)}K\ . \end{aligned} \tag{68}$$

As $\hat{\beta} \to \beta$ and $S_{\tilde{x}_y}^2 \to \sigma_x^2/f(\rho)$,

$$\sqrt{n}\,(\tilde{\beta} - \beta f(\rho)) \approx \sqrt{n}\,(\hat{\beta} - \beta + \beta K)f(\rho)\ . \tag{69}$$

References

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

[2] 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.

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

[4] 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.

[5] 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.

[6] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1*, 2nd edition. Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

[7] Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics, Volume 1*, 4th edition. Charles Griffin, London.

[8] 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.

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

[10] 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.

[11] 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.

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