Schmid:

Estimation of a Linear Model under Microaggregation by Individual Ranking

Sonderforschungsbereich 386, Paper 453 (2005)

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

Estimation of a Linear Model under Microaggregation by Individual Ranking

Matthias Schmid

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

Abstract

Microaggregation by individual ranking is one of the most commonly applied disclosure control techniques for continuous microdata. The paper studies the effect of microaggregation by individual ranking on the least squares estimation of a multiple linear regression model in continuous variables. It is shown that the naive parameter estimates are asymptotically unbiased. Moreover, the naive least squares estimates asymptotically have the same variances as the least squares estimates based on the original (non-aggregated) data. Thus, asymptotically, microaggregation by individual ranking does not induce any efficiency loss on the least squares estimation of a multiple linear regression model.

Keywords: Asymptotic variance, consistent estimation, disclosure control, individual ranking, linear model, microaggregation.

1 Introduction

The development of empirical research and the growing capacity of modern computer systems have lead to an increasing demand from researchers for access to microdata. Statistical offices and other data collecting institutions are therefore faced with the problem of providing statistically useful data sets that also comply with confidentiality requirements. One method to solve this problem is the application of masking procedures to data sets. The masking procedure itself is communicated to the researcher. Thus *anonymized* data sets with a low disclosure risk are created. However, masking a data set

usually implies that statistical analyses based on the masked data are less efficient or even biased. Statistical research is thus confronted with the problem of investigating the impact of masking techniques on parameter estimation, hypothesis testing, etc.

Over the last years, a wide variety of masking techniques has been developed, see Doyle et al. (2001), Willenborg and de Waal (2001), or Domingo-Ferrer (2002) for a general survey. The present paper deals with microaggregation, a very promising masking technique for continuous data (Anwar (1993), Defays and Nanopoulos (1993), Domingo-Ferrer and Mateo-Sanz (2002)). The basic principle of microaggregation is to subdivide a data set into small groups and to replace the original data values by their corresponding group means. There are various microaggregation techniques, which mainly differ in how the grouping of the data is done.

In the literature, many suggestions have been made on how to form the groups (see Domingo-Ferrer and Mateo-Sanz (2002)). To reduce the information loss imposed by microaggregation, only those data values which are "similar" to each other should be grouped (see Feige and Watts (1972)). The most commonly applied microaggregation techniques are microaggregation by single-axis sorting (SAS), multivariate microaggregation (MM), and microaggregation by individual ranking (IR). SAS uses a so-called *sorting variable* (e.g. a particular variable in the data set or the first principal component projection of the records) to determine the similarity of data values. MM uses a multivariate distance criterion (such as the Euclidean distance) to form the groups, while IR microaggregates all variables of a data set separately (see Section 2 for details). Analyses of selected data sets (Mateo-Sanz and Domingo-Ferrer (1998), Domingo-Ferrer and Mateo-Sanz (2001), Domingo-Ferrer and Torra (2001)) as well as an extensive simulation study performed by Schmid and Schneeweiss (2005) have shown that the efficiency loss induced by IR is relatively small if compared to other microaggregation techniques. However, several authors have pointed out that the disclosure risk of a data set remains relatively high if IR has been applied to it (see, e.g., Winkler (2002)).

In the present paper, the focus is on the effect of IR on the estimation of a multiple linear regression model. It is shown *analytically* that a linear regression in continuous variables can be consistently estimated by the naive least squares (LS) estimators if the data set has been anonymized by means of the

IR method. Thus the simulation results of Schmid and Schneeweiss (2005) are confirmed. Moreover, it is shown that the LS estimators based on the aggregated data are asymptotically as efficient as the LS estimators based on the non-aggregated (original) data. Thus, if the sample size is large, standard linear model techniques can be applied to the microaggregated data without leading to a severe bias or loss of efficiency.

The lemmas and theorems of the present paper partly rely on the results of Schmid et al. (2005a,b), who investigated the effect of microaggregation by a sorting variable on a simple linear regression model.

Section 2 starts with a description of the IR method. In Section 3 the effect of IR on the consistency of the LS estimation of a multiple linear regression model is investigated. Section 4 deals with the asymptotic variances of the naive LS estimators. Section 5 contains a simulation study on the results derived in Sections 3 and 4. In Section 6 the results are applied to the 2003 Munich rent data. A summary of the results of the present paper is contained in Section 7. Proofs of lemmas are given in the appendix.

2 Individual Ranking

The individual ranking technique considered in this paper works as follows: First, a fixed group size A has to be chosen. Next, the data set is sorted by the first variable, and groups of successive A values are formed. The values of the first variable in each group are replaced by their corresponding group mean, while the values of the other variables in the data set are left unchanged. Then the same procedure is repeated for the second variable, and so on.

For example, if a data set consists of the two vectors $x = (2, 6, 8, 1, 4, 3)$ and $y = (4, 6, 9, 8, 2, 7)$, the first step of IR results in the sorted data set

x	1	2	3	4	6	8
y	8	4	7	2	6	9

,

where the columns of the original data set are ordered according to the values of x. In the second step of IR, with A chosen to be 3, the values of x are microaggregated:

x	2	2	2	6	6	6
y	8	4	7	2	6	9

.

The third step of IR results in the sorted data set

x	6	2	6	2	2	6
y	2	4	6	7	8	9

,

where the columns are ordered according to the values of y. Finally, in the fourth step of IR, again with A chosen to be 3, the values of y are microaggregated:

x	6	2	6	2	2	6
y	4	4	4	8	8	8

.

3 Consistent Estimation

In this section the effect of IR on the least squares estimation of the multiple linear regression model

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

is analyzed. Y denotes the response (or endogenous variable), while $X_1, \ldots, X_p$ denote the covariates (or exogenous variables). Y and $X_1, \ldots, X_p$ are assumed to be continuous random variables with means μ_y, $\mu_x := (\mu_{x_1}, \ldots, \mu_{x_p})'$ and variances σ_y^2, $\sigma_{x_1}^2, \ldots, \sigma_{x_p}^2$. The supports of $Y, X_1, \ldots, X_p$ are (possibly infinite) intervals. The random error ϵ is assumed to be independent of $(X_1, \ldots, X_p)$ with zero mean and variance σ_ϵ^2. The objective is to estimate the parameter vector $(\beta_0, \beta_1, \ldots, \beta_p)'$ and the residual variance σ_ϵ^2 from an i.i.d. sample $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. Let $y := (y_1, \ldots, y_n)'$ and $x_j := (x_{1j}, \ldots, x_{nj})'$, $j = 1, \ldots, p$, contain the data values. The vectors containing the aggregated data are denoted by $\tilde{y}$ and $\tilde{x}_1, \ldots, \tilde{x}_p$. For simplicity, it is assumed throughout that n is a multiple of A. Define

$$\beta := (\beta_1, \ldots, \beta_p)' , \tag{2}$$

$$X := (x_1, \ldots, x_n) , \tag{3}$$

$$\tilde{X} := (\tilde{x}_1, \ldots, \tilde{x}_n) . \tag{4}$$

It is now shown that the least squares estimator

$$\begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_p \end{pmatrix} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta} \end{pmatrix} := \big((\mathbf{1}\tilde{X})'(\mathbf{1}\tilde{X})\big)^{-1}\big(\mathbf{1}\tilde{X}\big)'\tilde{y} \tag{5}$$

is a consistent estimator of $(\beta_0, \beta')'$. First consider $\tilde{\beta}$, which is the estimator of the slope parameter vector β. Without loss of generality assume that the vectors $\tilde{y}, \tilde{x}_1, \ldots, \tilde{x}_p$ are centered around their means $\bar{\tilde{y}}, \bar{\tilde{x}}_1, \ldots, \bar{\tilde{x}}_p$. Note that $\bar{\tilde{y}} = \bar{y}, \bar{\tilde{x}}_1 = \bar{x}_1, \ldots, \bar{\tilde{x}}_p = \bar{x}_p$. By definition,

$$\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = \begin{pmatrix} \tilde{S}^2_{x_1} & \tilde{S}_{x_1x_2} & \ldots & \tilde{S}_{x_1x_p} \\ \vdots & \vdots & & \vdots \\ \tilde{S}_{x_px_1} & \tilde{S}_{x_px_2} & \ldots & \tilde{S}^2_{x_p} \end{pmatrix} \cdot \begin{pmatrix} \tilde{S}_{x_1y} \\ \vdots \\ \tilde{S}_{x_py} \end{pmatrix} =: \tilde{S}_{xx} \cdot \tilde{S}_{xy} \ , \tag{6}$$

where $\tilde{S}^2_{x_1}, \ldots, \tilde{S}^2_{x_p}$ are the empirical variances of $\tilde{x}_1, \ldots, \tilde{x}_p$, respectively. The expressions $\tilde{S}_{x_jx_k}$, $j, k = 1, \ldots, p$, $j \neq k$, are the empirical covariances of $\tilde{x}_j$ and $\tilde{x}_k$. Analogously, $\tilde{S}_{x_jy}$, $j = 1, \ldots, p$, are the empirical covariances of $\tilde{x}_j$ and $\tilde{y}$. The probability limit of $\tilde{\beta}$ can be obtained by deriving the probability limits of the variances and covariances in equation (6).

Consider first the empirical variances $\tilde{S}^2_{x_1}, \ldots, \tilde{S}^2_{x_p}$. As will be shown, $\tilde{S}^2_{x_1}, \ldots, \tilde{S}^2_{x_p}$ are consistent estimators of the true variances $\sigma^2_{x_1}, \ldots, \sigma^2_{x_p}$.

Lemma 1. *Assume Z to be a continuous random variable with variance σ^2_z whose support is a (possibly infinite) interval. Denote by $\tilde{S}^2_z$ the empirical variance of the aggregated data values $\tilde{z}_1, \ldots, \tilde{z}_n$ based on an i.i.d. sample $(z_1, \ldots, z_n)$. Then*

$$\tilde{S}^2_z \xrightarrow{p} \sigma^2_z \ . \tag{7}$$

Proof. See Schmid et al. (2005a). □

It follows from Lemma 1 that $\tilde{S}^2_{x_1}, \ldots, \tilde{S}^2_{x_p}$ converge in probability to $\sigma^2_{x_1}, \ldots, \sigma^2_{x_p}$. Next, consider the covariances $\tilde{S}_{x_jx_k}$ and $\tilde{S}_{x_jy}$. As will be shown,

$\tilde{S}_{x_j x_k}$ and $\tilde{S}_{x_j y}$ are consistent estimators of the true covariances $\sigma_{x_j x_k}$ and $\sigma_{x_j y}$.

Lemma 2. *Assume Z_1 and Z_2 to be continuous random variables with variances $\sigma_{z_1}^2$ and $\sigma_{z_2}^2$. Assume the supports of Z_1 and Z_2 to be (possibly infinite) intervals. Denote by $\sigma_{z_1 z_2}$ the covariance of Z_1 and Z_2. Further denote by $\tilde{S}_{z_1 z_2}$ the empirical covariance of the aggregated data values $\tilde{z}_1 := (\tilde{z}_{i1}, \ldots, \tilde{z}_{n1})'$ and $\tilde{z}_2 := (\tilde{z}_{i2}, \ldots, \tilde{z}_{n2})'$ based on an i.i.d. sample (z_{i1}, z_{i2}), $i = 1, \ldots, n$. Then $\tilde{S}_{z_1 z_2}$ converges in probability to $\sigma_{z_1 z_2}$.*

Proof. See appendix. □

The consistency of $(\tilde{\beta}_0, \tilde{\beta}')'$ is stated in the following theorem:

Theorem 1. *The least squares estimator $(\tilde{\beta}_0, \tilde{\beta}')'$ is a consistent estimator of the true parameter vector $(\beta_0, \beta')'$.*

Proof. The consistency of $\tilde{\beta}$ follows, due to equation (6), from Lemma 1 and Lemma 2. Moreover,

$$\tilde{\beta}_0 = \tilde{\bar{y}} - \tilde{\beta}' \begin{pmatrix} \tilde{\bar{x}}_1 \\ \vdots \\ \tilde{\bar{x}}_p \end{pmatrix} = \bar{y} - \tilde{\beta}' \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} \xrightarrow{p} \mu_y - \beta' \mu_x = \beta_0 \ . \tag{8}$$

Thus $\tilde{\beta}_0$ is a consistent estimator of β_0. □

Denote by $\tilde{\sigma}_\epsilon^2$ the least squares estimator of the residual variance σ_ϵ^2. By definition,

$$\tilde{\sigma}_\epsilon^2 = \tilde{S}_y^2 - \tilde{\beta}' \tilde{S}_{xx} \tilde{\beta} \ , \tag{9}$$

where $\tilde{S}_y^2$ is the empirical variance of $\tilde{y}$. From Lemmas 1 and 2 and from Theorem 1 it follows that $\text{plim}_{n\to\infty} \tilde{\sigma}_\epsilon^2 = \sigma_\epsilon^2$. Thus $\tilde{\sigma}_\epsilon^2$ is a consistent estimator of σ_ϵ^2.

4 Asymptotic Variance of the Least Squares Estimator

In this section the asymptotic variance of the least squares estimator $(\tilde{\beta}_0, \tilde{\beta}')'$ is derived. To do so, stronger assumptions than in Section 3 are required: $Y, X_1, \ldots, X_p$ are now assumed to be jointly normally distributed with mean $\mu := (\mu_y, \mu_x')'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \Sigma_{xy}' \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} := \begin{pmatrix} \sigma_y^2 & \sigma_{x_1 y} & \ldots & \ldots & \sigma_{x_p y} \\ \sigma_{x_1 y} & \sigma_{x_1}^2 & \sigma_{x_1 x_2} & \ldots & \sigma_{x_1 x_p} \\ \vdots & \vdots & \vdots & & \vdots \\ \sigma_{x_p y} & \sigma_{x_1 x_p} & \sigma_{x_2 x_p} & \ldots & \sigma_{x_p}^2 \end{pmatrix} . \tag{10}$$

Some additional notation is required:

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

The derivation of the asymptotic variance of $(\tilde{\beta}_0, \tilde{\beta}')'$ is based on the following lemma, which uses the notations and assumptions of Lemma 2:

Lemma 3. *Assume Z_1 and Z_2 to be jointly normally distributed. Then*

a) $\tilde{S}_{z_1}^2$ and $\tilde{S}_{z_2}^2$ are asymptotically equivalent to $S_{z_1}^2$ and $S_{z_2}^2$, respectively.

b) $\tilde{S}_{z_1 z_2}$ is asymptotically equivalent to $S_{z_1 z_2}$.

Proof. a) See Schmid et al. (2005b). b) See appendix. □

The asymptotic variance of $(\tilde{\beta}_0, \tilde{\beta}')'$ is characterized by the following theorem:

Theorem 2. *Assume $Y, X_1, \ldots, X_p$ to be jointly normally distributed random variables. The asymptotic variance of the least squares estimator $(\tilde{\beta}_0, \tilde{\beta}')'$ computed from the microaggregated data is equal to the asymptotic variance of the least squares estimator $(\hat{\beta}_0, \hat{\beta}')'$ computed from the non-aggregated data.*

Proof. Denote by S_{xx} the empirical covariance matrix of $x_1, \ldots, x_p$ and by S_{xy} the empirical covariance vector of $x_1, \ldots, x_p$ and y. Let $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_p)'$. By (6) and (8), $(\tilde{\beta}_0, \tilde{\beta}')'$ can be expressed as

$$(\tilde{\beta}_0, \tilde{\beta}')' = G(\tilde{S}_{xx}, \tilde{S}_{xy}, \bar{\tilde{x}}, \bar{\tilde{y}}) = G(\tilde{S}_{xx}, \tilde{S}_{xy}, \bar{x}, \bar{y}) \ , \tag{11}$$

where G is a continuously differentiable function. A similar relation with the same function G holds for $(\hat{\beta}_0, \hat{\beta}')'$:

$$(\hat{\beta}_0, \hat{\beta}')' = G(S_{xx}, S_{xy}, \bar{x}, \bar{y}) \ . \tag{12}$$

Now, by Lemma 3, $\tilde{S}_{xx} \sim S_{xx}$ and $\tilde{S}_{xy} \sim S_{xy}$. In addition, $\sqrt{n}(S_{xx} - \Sigma_{xx})$, $\sqrt{n}(S_{xy} - \Sigma_{xy})$, $\sqrt{n}(\bar{x}_1 - \mu_{x_1}), \ldots, \sqrt{n}(\bar{x}_p - \mu_{x_p})$, and $\sqrt{n}(\bar{y} - \mu_y)$ are asymptotically bounded. Therefore

$$G(S_{xx}, S_{xy}, \bar{x}, \bar{y}) \sim G(\tilde{S}_{xx}, \tilde{S}_{xy}, \bar{\tilde{x}}, \bar{\tilde{y}}) \ . \tag{13}$$

Thus $(\hat{\beta}_0, \hat{\beta}')'$ and $(\tilde{\beta}_0, \tilde{\beta}')'$ are asymptotically equivalent, which implies the theorem.

□

It follows from Theorem 2 that $(\tilde{\beta}_0, \tilde{\beta}')'$ and $(\hat{\beta}_0, \hat{\beta}')'$ are asymptotically equally efficient. Thus, asymptotically, microaggregation by individual ranking does not impose any efficiency loss on the least squares estimation of model (1). The asymptotic variance of $(\tilde{\beta}_0, \tilde{\beta}')'$ can be estimated by

$$\hat{\Sigma}_{\tilde{\beta}} := \widehat{\mathrm{var}}((\tilde{\beta}_0, \tilde{\beta}')') = \tilde{\sigma}_\epsilon^2((\mathbf{1}\tilde{X})'(\mathbf{1}\tilde{X}))^{-1} \ . \tag{14}$$

With the help of (14), asymptotic confidence intervals for $\beta_0, \beta_1, \ldots, \beta_p$ can be constructed:

$$CI_{\beta_j} = \left[\tilde{\beta}_j - z_{1-\alpha/2}\sqrt{\hat{\Sigma}_{\tilde{\beta}(jj)}}\,,\ \tilde{\beta}_j + z_{1-\alpha/2}\sqrt{\hat{\Sigma}_{\tilde{\beta}(jj)}} \right]\ ,\ j = 0, 1, \ldots, p\ , \quad (15)$$

where $\hat{\Sigma}_{\tilde{\beta}(jj)}$ is the j-th diagonal element of $\hat{\Sigma}_{\tilde{\beta}}$ and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

5 Simulations

In this section the asymptotic results of Sections 3 and 4 are investigated by means of a simulation study. Thus the finite sample behavior of the naive LS estimators can be studied under microaggregation. For the simulation study, a linear model with two standard normal covariates X_1 and X_2 was estimated from 500 independent samples. Each sample was microaggregated by means of the IR method before the linear model was estimated. The residual error variance σ_ϵ^2 was set to 9, the group size A was set to 3. Throughout this section, the slope parameter β_2 is kept fixed ($\beta_2 = 1$).

Fig. 1 shows the estimated bias of $\tilde{\beta}_1$ for different values of β_1 and various sample sizes n. The covariates X_1 and X_2 are uncorrelated. Apparently, if n is small and β_1 is positive, the naive least squares estimate of β_1 is negatively biased. For moderate sample sizes ($n = 51$, $n = 150$) the bias almost disappears. Fig. 2 shows the corresponding variances of $\tilde{\beta}_1$ (multiplied with $\sqrt{n}$). Apparently, for small n, the asymptotic variance of $\sqrt{n}\tilde{\beta}_1$ differs from the true variance of $\sqrt{n}\tilde{\beta}_1$, but for moderate values of n ($n = 99$, $n = 150$) the approximation is very good. Figs. 3 and 4 show the bias and the variance of $\tilde{\beta}_1$ when X_1 and X_2 are moderately correlated ($\mathrm{cor}(X_1, X_2) = 0.5$). Apparently, the bias and variance of $\tilde{\beta}_1$ are larger than in Figs. 1 and 2 (where X_1 and X_2 are uncorrelated). Again, if the sample size is small, the true parameters differ from their asymptotic counterparts. As n increases, the differences disappear. If X_1 and X_2 are highly correlated ($\mathrm{cor}(X_1, X_2) = 0.8$), the bias and variance of $\tilde{\beta}_1$ increase even further (see Figs. 5 and 6). However, the approximations derived in Sections 3 and 4 are good when the sample size is higher than $n = 99$.

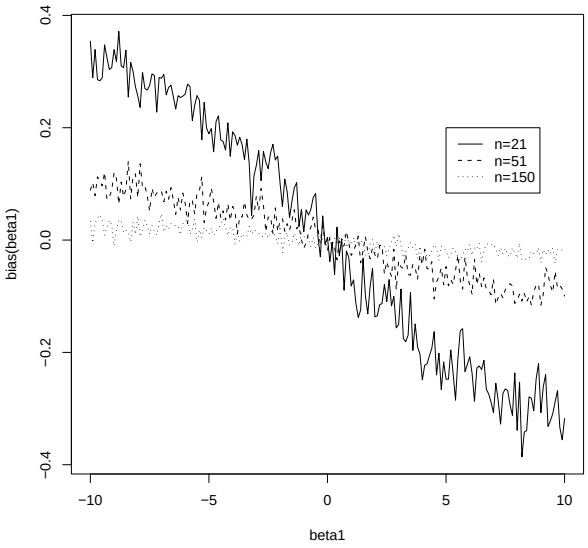

Figure 1: Bias of $\tilde{\beta}_1$ ($\text{cor}(X_1, X_2) = 0$)

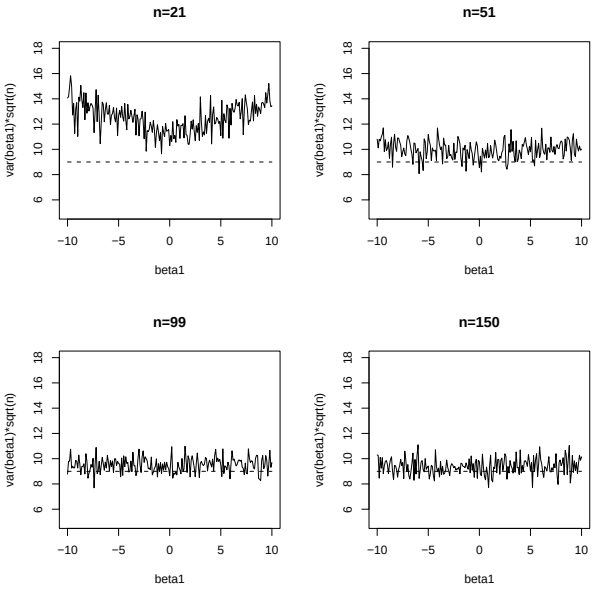

Figure 2: Variance of $\sqrt{n}\tilde{\beta}_1$ ($\text{cor}(X_1, X_2) = 0$), dashed line = asymptotic variance of $\sqrt{n}\tilde{\beta}_1$

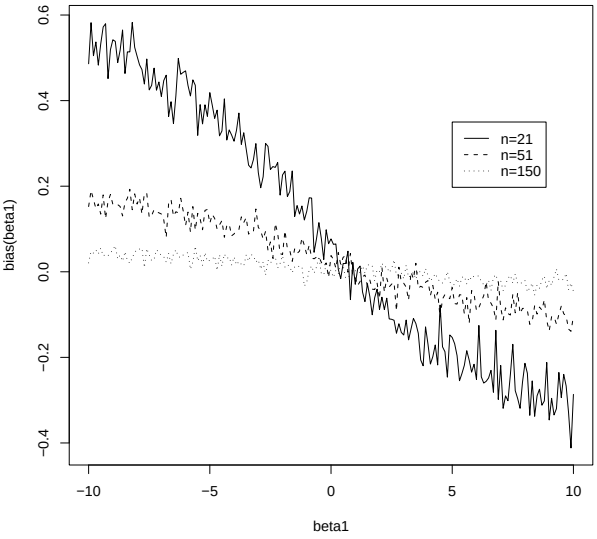

Figure 3: Bias of $\tilde{\beta}_1$ ($\text{cor}(X_1, X_2) = 0.5$)

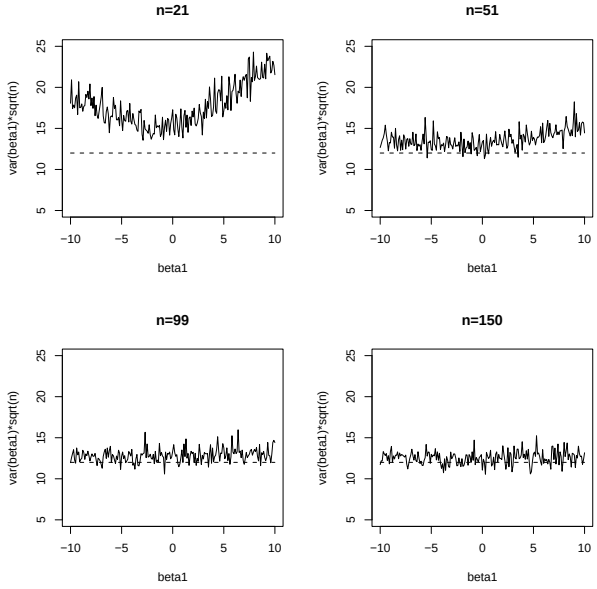

Figure 4: Variance of $\sqrt{n}\tilde{\beta}_1$ ($\text{cor}(X_1, X_2) = 0.5$), dashed line = asymptotic variance of $\sqrt{n}\tilde{\beta}_1$

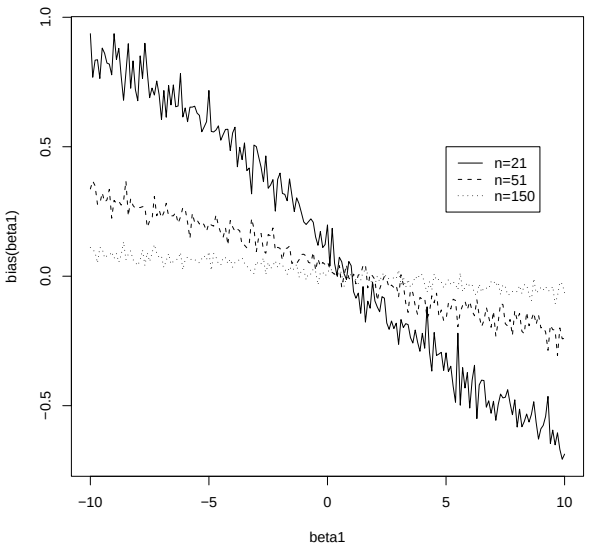

Figure 5: Bias of $\tilde{\beta}_1$ ($\text{cor}(X_1, X_2) = 0.8$)

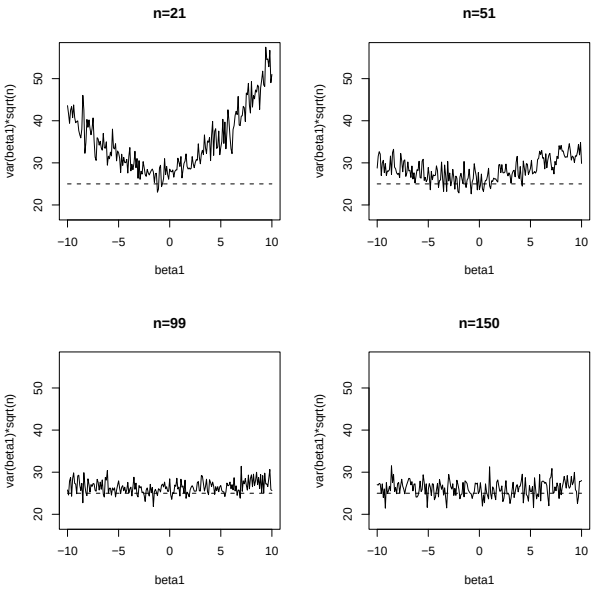

Figure 6: Variance of $\sqrt{n}\tilde{\beta}_1$ ($\text{cor}(X_1, X_2) = 0.8$), dashed line = asymptotic variance of $\sqrt{n}\tilde{\beta}_1$

6 Munich Rent Data

In this section the results derived in Sections 3 and 4 are applied to the 2003 Munich rent data (`http://www.statistik.lmu.de/service/datenarchiv/miete/miete03_e.html`). The data set contains 2053 households interviewed for the 2003 Munich rent standard. As it is publicly available, the *original* parameter estimates can be computed, and the impact of microaggregation on a linear regression can be studied directly. In the following, the relationship between the monthly net rent of the households in EUR (`nr`, dependent variable), the floor space in m^2 (`fs`, independent variable), and the year of construction of the buildings (`yc`, independent variable) is analyzed. First, a linear model based on the original (non-aggregated) data is estimated. The resulting estimates are then compared to the linear model estimates based on the data set which has been aggregated by means of the IR method (with group size $A = 3$).

The results of the two analyses are contained in Tables 1 and 2. As expected, microaggregation by individual ranking has almost no influence on the parameter estimates. The residual standard error estimates (167.0749 in case of the non-aggregated data, 166.3248 in case of the aggregated data) are also very similar.

Moreover, Tables 1 and 2 show that microaggregation by individual ranking has almost no influence on the estimated standard errors of the coefficients of `fs` and `yc`. This is remarkable because `nr`, `fs`, and `yc` do not follow a normal distribution (compare Figs. 7 to 9), which means that one important assumption made in the proof of Theorem 2 does not hold for the Munich rent data. The estimation procedure thus seems to be robust against violations of the normality assumption.

	Estimate	Std. Error	t-value	p-value
Intercept	-3716.1958	298.4688	-12.45	0.0000
fs	7.2804	0.1496	48.67	0.0000
yc	1.9304	0.1513	12.76	0.0000

Table 1: Regression of `nr` on `fs` and `yc` (non-aggregated data)

	Estimate	Std. Error	t-value	p-value
Intercept	-3712.0200	296.9041	-12.50	0.0000
fs	7.3017	0.1488	49.07	0.0000
yc	1.9275	0.1505	12.81	0.0000

Table 2: Regression of `nr` on `fs` and `yc` (aggregated data)

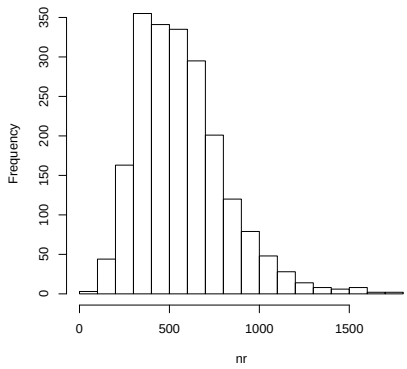

Figure 7: Histogram of `nr`

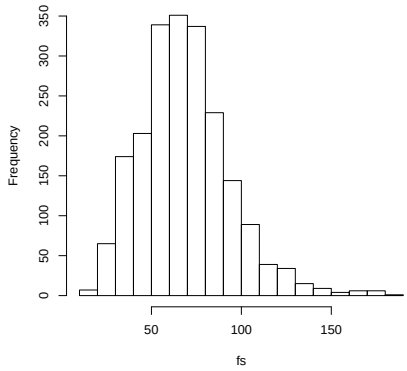

Figure 8: Histogram of `fs`

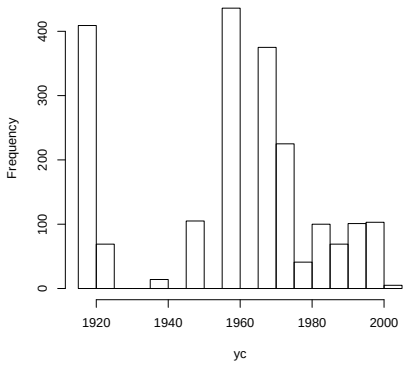

Figure 9: Histogram of `yc`

7 Conclusion

Microaggregation based on individual ranking is one of the most commonly applied disclosure control techniques for continuous data. The present paper deals with the impact of IR on statistical analysis. It shows that if a multiple linear regression model in continuous variables is estimated from the aggregated data, naive least squares estimates are asymptotically unbiased. Moreover, if the dependent variable and the independent variables are jointly normally distributed, parameter estimates based on the aggregated data are asymptotically as efficient as the estimates based on the non-aggregated data. Thus the efficiency loss induced by IR is negligible if the sample size is high.

The simulation study carried out in Section 5 shows that the finite bias of the parameter estimates is close to 0 if the sample size is moderately high ($n = 51$). For small sample sizes ($n = 21$), the naive LS estimates are biased. Similarly, if the sample size is moderately high ($n = 99$), the finite sample variances of the parameter estimates are close to their asymptotic variances (which are the variances of the parameter estimates based on the non-aggregated data). This implies that the efficiency loss induced by IR is relatively small. For small sample sizes ($n = 21$), the sample variances of the LS estimates differ from their asymptotic counterparts.

The analysis of the Munich rent data in Section 6 shows that even if the variables are not normally distributed, variance estimates of the model co-

efficients do not differ notably from the variance estimates based on the non-aggregated data. Thus the least squares estimators seem to be robust against violations of the normality assumption.

In summary, microaggregation by individual ranking asymptotically has no influence on the least squares estimation of linear regression in continuous variables. This is not generally true for other microaggregation techniques (see Schmid et al. (2005a,b)) or for masking procedures such as the addition of random noise (see Brand (2000)). Thus, although IR implies a relatively high disclosure risk, its analytical properties are far superior to those of many other anonymization techniques.

Appendix

Proof of Lemma 2:

Each data value z_{i1}, $i = 1, \ldots, n$, can be written as

$$z_{i1} = \tilde{z}_{i1} + \eta_{i1} \ , \tag{16}$$

where η_{i1} is the difference between the original data value z_{i1} and the aggregated data value $\tilde{z}_{i1}$. Now the empirical variance S^2_{z1} of $z_1 := (z_{11}, \ldots, z_{n1})'$ can be decomposed as follows:

$$S^2_{z_1} = \tilde{S}^2_{z_1} + S^2_{\eta_1} \ , \tag{17}$$

where $\tilde{S}^2_{z_1}$ is the between-groups variance of z_1 and $S^2_{\eta_1}$ is the within-groups variance of z_1. In Schmid et al. (2005a) it was shown that $S^2_{\eta_1}$ converges to 0 in probability as $n \to \infty$. Thus, $\text{plim}_{n\to\infty}\tilde{S}^2_{z_1} = \sigma^2_{z_1}$. Analogously, $z_{i2} = \tilde{z}_{i2} + \eta_{i2}$, $i = 1, \ldots, n$, and $\text{plim}_{n\to\infty}S^2_{\eta_2} = 0$.

Denote the empirical covariance of the non-aggregated data vectors z_1 and z_2 by $S_{z_1 z_2}$. Define $z^*_{1i} := z_{1i} - \bar{z}_1$ and $z^*_{2i} := z_{2i} - \bar{z}_2$. Now, $\tilde{z}_{1i} - \bar{\tilde{z}}_1 = z^*_{1i} - \eta_{1i}$

and $\tilde{z}_{2i} - \bar{\tilde{z}}_2 = z^*_{2i} - \eta_{2i}$ (as $\bar{\eta}_1 = \bar{\eta}_2 = 0$). One obtains

$$\begin{aligned} |\tilde{S}_{z_1 z_2} - S_{z_1 z_2}| &= \frac{1}{n}\left|\sum_{i=1}^{n}(z^*_{i1} - \eta_{i1})(z^*_{i2} - \eta_{i2}) - \sum_{i=1}^{n} z^*_{i1} z^*_{i2}\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^{n} z^*_{i1}\eta_{i2}\right| + \left|\frac{1}{n}\sum_{i=1}^{n} z^*_{i2}\eta_{i1}\right| + \left|\frac{1}{n}\sum_{i=1}^{n} \eta_{i1}\eta_{i2}\right| \\ &\leq \sqrt{S^2_{z_1} S^2_{\eta_2}} + \sqrt{S^2_{z_2} S^2_{\eta_1}} + \sqrt{S^2_{\eta_1} S^2_{\eta_2}} \xrightarrow{p} 0 \,. \end{aligned} \tag{18}$$

Proof of Lemma 3b):

Analogously to the proof of Lemma 2, $S^2_{z_1}$ and $S^2_{z_2}$ can be decomposed into $\tilde{S}^2_{z_1} + S^2_{\eta_1}$ and $\tilde{S}^2_{z_2} + S^2_{\eta_2}$, respectively. From Lemma 3a) it follows that $\text{plim}_{n\to\infty}\sqrt{n}S^2_{\eta_1} = 0$ and $\text{plim}_{n\to\infty}\sqrt{n}S^2_{\eta_2} = 0$. Hence by (18)

$$\sqrt{n}|\tilde{S}_{z_1 z_2} - S_{z_1 z_2}| \leq \sqrt{S^2_{z_1}}\sqrt{n}\sqrt{S^2_{\eta_2}} + \sqrt{S^2_{z_2}}\sqrt{n}\sqrt{S^2_{\eta_1}} + \sqrt{n}\sqrt{S^2_{\eta_1}}\sqrt{S^2_{\eta_2}} \xrightarrow{p} 0 \,. \tag{19}$$

Acknowledgements

I thank Hans Schneeweiss for very helpful discussions and comments.

References

Anwar, M. N. (1993): "Micro-Aggregation - The Small Aggregates Method," Internal report, Eurostat, Luxembourg.

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

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

Domingo-Ferrer, J. (2002): *Inference Control in Statistical Databases*. New York: Springer.

Domingo-Ferrer, J. and J. M. Mateo-Sanz (2001): "An Empirical Comparison of SDC Methods for Continuous Microdata in terms of Information Loss and Re-Identification Risk," Second Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, Macedonia.

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

Domingo-Ferrer, J. and V. Torra (2001): "A Quantitative Comparison of Disclosure Control Methods for Microdata". In *Confidentiality, Disclosure, and Data Access*, ed. by P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz. Amsterdam: North-Holland, 111-133.

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

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

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

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

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

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

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

Winkler, W. E. (2002): "Single-Ranking Micro-Aggregation and Re-identification," Statistical Research Division report RR 2002/08, U.S. Bureau of the Census, Washington.