(Generalized) Linear Regression on Microaggregated Data – From Nuisance Parameter Optimization to Partial Identification

Paul Fink PAUL.FINK@STAT.UNI-MUENCHEN.DE
Thomas Augustin AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics, Ludwig-Maximilians-Universität München (LMU Munich)
Munich (Germany)

Abstract

Protecting sensitive micro data prior to publishing or passing the data itself on is a crucial aspect: A trade-off between sufficient disclosure control and analyzability needs to be found. This paper presents a starting point to evaluate the effect of k-anonymity microaggregated data in (generalized) linear regression. Taking a rigorous imprecision perspective, microaggregated data are understood inducing a set $\mathbb{X}$ of potentially true data. Based on this representation two conceptually different approaches deriving estimations from the *ideal* likelihood are discussed. The first one picks a single element of $\mathbb{X}$, for instance by naively treating the microaggregated data as true ones or by introducing a maximax approach taking the elements of $\mathbb{X}$ as nuisance parameters to be optimized. The second one seeks, in the spirit of Partial Identification, the set of all maximum likelihood estimators compatible with the elements of $\mathbb{X}$, thus creating cautious estimators. As the simulation study corroborates, the obtained sets of estimators of the latter approach are still precise enough to be practically relevant.

Keywords: maximum likelihood estimation; generalized linear regression; microaggregation; anonymization; partial identification.

1. Introduction

In recent years, more data are made available, for instance in the context of websites for marketing purposes or also by institutions of Official Statistics[1]. These micro data usually contain sensitive information of the units involved. As the combination of different disjoint data sources requires lesser effort now, data obtained in one isolated context may be not revealing, however, the join of multiple data sources may make the unit identifiable. Therefore, powerful anonymization techniques for disclosure control, inducing an information reduction, are essential to protect the privacy and strengthen the collected data quality. Scientific researchers have a diametrically opposite aim: They desire a deeper understanding of the unit's action and/or the social or economic processes involved, hence the availability of minute information about the units under consideration is essential. As data sources they may rely either on self-collected data or on such from a different source, for instance Official Statistics or private companies. Henceforth a trade-off between granting sufficient privacy while still ensuring data utility is required. In cases when the anonymization technique does not provide sufficient information after its deployment, the availability of the data itself is reduced to absurdity as the data provision might be scrapped entirely.

A well known concept in this setting is the so-called *k-anonymity* proposed by Sweeney (2002): It guarantees that each value of each anonymized variable occurs at least k times. Hence, even if

1. The European Statistics Code of Practice explicitly encourages to make the collected data publicly available (cf. Eurostat and European Statistical System (2011, principle 15))

attackers know identifying aspects of one record within the micro data, they are unable to deduce the actual value of the sensitive variable in question. One specific concept ensuring k-anonymity is *microaggregation*, which belongs to the family of perturbative methods in statistical disclosure control (e.g. Willenborg and de Waal (2001)), replacing individual records by a representative substitute, e.g. a group average. Initially developed to deal with just continuous variables, microaggregation is nowadays applicable to any measurement scale of the variable(s) to anonymize. However, in this paper the focus is on continuous variables. Microaggregation by a given technique is herein understood as a mapping m, which by design maps several situations of the values to be microaggregated $\boldsymbol{x}$ onto the same microaggregation result in the image $\tilde{\boldsymbol{x}}$. However, when analyzing microaggregated data, one is interested in findings on the original data and therefore in the reverse mapping which induces a set of compatible underlying data situations

$$\mathbb{X}(\tilde{\boldsymbol{x}}) = \{\boldsymbol{x} \mid m(\boldsymbol{x}) = \tilde{\boldsymbol{x}}\} \ .$$

In literature, microaggregation techniques have been evaluated mostly in the light of disclosure control, but few on the potential for analyses: To which extent are structures in the original data inferable by looking at the anonymized data? This paper deals with linear regression, a standard statistical model strategy which is commonly employed in econometrics, biometrics and social sciences. As basis for further research it is considered in the broader framework of generalized linear regression. In the standard case for precise observations the estimators for the structural parameters are obtained when maximizing the likelihood with respect to those. In order to obtain concise estimators in the case of microaggregated data, one could just ignore the nature of the data and use them as-is or employ a maximax-like approach by maximizing the likelihood with respect to the structure parameters and all compatible underlying data situations, suggested herein as first way to proceed. By taking only such an optimistic instantiation of $\mathbb{X}$ the introduced imprecision is not taken seriously. Therefore, in the spirit of Partial Identification (Manski, 2003) a more cautious estimation approach is introduced by reporting the set of all maximum likelihood estimators based on the elements of $\mathbb{X}$. All findings are derived in case of a generic microaggregation and thus are suitable for any microaggregation technique. However, it is also demonstrated how the estimation is further improved when considering the specific masking microaggregation technique.

This paper is structured as follows: In section 2 a short overview of microaggregation as anonymization technique and of generalized linear regression is given. In section 3 approaches to obtain concise estimators are presented, while in section 4 the partial identification view is taken resulting in sets-valued estimators. A simulation study in section 5 evaluates the theoretically obtained results. The paper concludes with some remarks and an outlook for further research in section 6.

2. Microaggregation and (Generalized) Linear Regression

2.1 Basics of Microaggregation

The general idea of microaggregation is to replace the individual records by a representative substitute of at least k individuals, in turn microaggregation in this sense then satisfies k-anonymity. Any microaggregation technique may be represented as a two-step process.

Grouping: The individual records of the micro data are partitioned into clusters in a certain way such that records within a cluster are similar and each cluster contains at least $k \geq 3$ records.

Aggregation: Each individual record within a cluster is replaced by the cluster's characteristic value, e.g. mean or median.

The choice on how to define similarity of observations and how to deal with multiple variables allows for a variety of actual techniques. As minimal requirement the previously mentioned concept of k-anonymity (Sweeney, 2002) needs to be fulfilled, guaranteeing that each value of each anonymized variable occurs at least k times. In the simplest case of a single variable, neighboring observations are grouped together and their values are replaced by their group mean. Thus without knowledge about the original data, the membership to the employed groups in the first microaggregation step are deducible as same values in the microaggregated data indicate membership to the same group.

Microaggregation techniques relying on a sorting of variable(s) include *Single-axis Sorting*, where the data are globally ordered according to single (external) sorting variable, and *Individual Ranking*, in which for every variable to be microaggregated an independent Single-Axis sorting is applied according to the ranks of itself. From the perspective of disclosure control it is seen critically that regions for the underlying true records are deducible for both Individual Ranking and Single-axis Sorting in cases when one of the variables to anonymize acts as the sorting variable.[2] From the analyst's view those regions are exploitable when estimating statistical models as will be seen in the following sections. Other microaggregation techniques do not rely on the concept of an underlying sorting variable, but employ directly a multivariate clustering, e.g. *Maximum Distance to Average Vector* (MDAV) by Domingo-Ferrer and Mateo-Sanz (2002), also providing natural regions.

As desired, the grouping and aggregation introduce imprecision, which means that several different data sets of the underlying true data will lead to the same microaggregated data set. In case of inference on structures in the underlying true data one needs to account for this imprecision. In this paper two conceptually different views are presented: The first, often implicitly found in the literature without embedding into a formal framework, will prove characterizable as a *maximax*-like approach, where the most plausible data situation(s) in the light of the estimation function are used to obtain the parameter estimate(s). The second is the Partial Identification view, where instead of an optimistic estimator, a set-valued one is reported, reflecting the imprecision in the input data more accurately.

2.2 (Generalized) Linear Regression in a Nutshell

The two views are presented for the case of classical linear regression, in order to prepare its extension in the formulation of the generalized linear regression framework, which uses a maximum likelihood approach for parameter estimation instead of the ordinary least squares. The basic settings of the likelihood approach are briefly sketched now[3]: The aim of (generalized) linear regression is to model the dependency of p independent covariates $\boldsymbol{X} = (X_1, \ldots, X_q, \ldots, X_p)$ on a response variable Y, without claiming a causal relation in either direction. For each unit $i = 1, \ldots, n$ the response y_i and the covariates $\boldsymbol{x}_i$ are observed, densely written as $\boldsymbol{y} = (y_1, \ldots, y_i, \ldots, y_n)^T$ and $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_i^T, \ldots, \boldsymbol{x}_n^T)^T$. The dependency is modeled by means of a linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = (1, \boldsymbol{x}_i)\boldsymbol{\beta}$, which is in the context of generalized regression transformed by a so-called *response* function[4] h to model the conditional expectation $\mathbb{E}(Y|\boldsymbol{X})$, where

2. Rosemann et al. (2004) argues that in some practically relevant contexts it is negligible.
3. For further details see Fahrmeir et al. (2013, p. 301ff.)
4. h^{-1} is the so-called *link function*.

the form of the conditional distribution is to be specified as a certain one from the exponential family. Common choices lead to well-known models, e.g. a normal distribution for classical linear regression or a Bernoulli distribution for logistic regression; in the general formulation several models are dealt with.

The parameters of interest $\boldsymbol{\beta}$ are estimated by maximizing the (log-)likelihood induced by the modeling assumptions or equivalently by solving the equation system when setting the *score function* (derivate of the log-likelihood with respect to the parameters of interest) to zero. The score function in classical linear regression for β_q is obtainable to

$$s(\beta_q) = \frac{1}{\phi} \sum_{i=1}^{n} x_{iq} \Big(y_i - (1, \boldsymbol{x}_i)\boldsymbol{\beta} \Big) \,, \tag{1}$$

where in the case $q = 0$ the value of x_{i0} is set to one.[5]

For a start only microaggregation of all of the covariates is covered in this work; a situation common in social sciences, when the effect of several covariates, requiring anonymization, on a response variable, for which anonymization is not required, is to be estimated. An extension of this situation to mixtures of aggregation techniques employed for covariates of different scale is subject to further research. It is believed that the findings presented herein may be adapted straightforwardly to the case when only some covariates are microaggregated and the others are left as-is. Considering a microaggregated response variable appears more difficult as conditional independence of the now microaggregated y_i's in the likelihood does no longer hold. Moreover, it is believed that the actual choice of microaggregation technique has a more severe impact on the quality of the estimation.

2.3 Common Structural Implications of Microaggregation

Before discussing microaggregation in the context of (generalized) linear regression, the structure of microaggregation is recalled. To draw the distinction between the original values and its microaggregated counterparts, as briefly sketched in the introduction, the first are denoted by $\boldsymbol{x}$, whereas for the microaggregated values a tilde is placed above: $\tilde{\boldsymbol{x}}$. A technique suitable for k-anonymity is assumed with k as the fixed group size. For simplicity reasons within this paper it is further assumed that the number of observation n is a multiple of k. Nonetheless, the proposed approach may be straightforwardly generalized to the case when k is only a minimal group size. As a result there are $G = n/k$ distinct groups for each variable under microaggregation. As the microaggregation process involves grouping and averaging, depending on the actual technique chosen even per variable, the notation for generic microaggregation described herein is severely affected. In order to index an observation, a two level indexing in the superscript is utilized: The first place is taken by the membership in a specific group and in the second place the index within this group is given. To denote the grouping according to a specific microaggregated covariate, the group label is further indexed by it. This means, $\tilde{x}_q^{g_r,j}$ corresponds to the microaggregated value of the q^{th} covariate for the j^{th} observation in the g^{th} group, when grouping is induced by the microaggregation of the r^{th} covariate. Please recall that $\tilde{x}_q^{g_q,j}$ has the same value for $j = 1, \ldots, k$.

In the above notation the original values are representable by adding a further parameter level:

$$x_q^{g_r,j} = \tilde{x}_q^{g_r,j} + \Delta_q^{g_r,j} \,, \tag{2}$$

5. The estimation equation for generalized linear regression with canonical link function takes a similar form, only replacing $(1, \boldsymbol{x}_i)$ by $h(1, \boldsymbol{x}_i)$

where $\Delta_q^{g_r,j}$ is the corresponding deviation of the individual record from its corresponding group mean. By just looking at the microaggregated values, one is able to deduce the group membership used in the microaggregation's aggregation step. As the mean value within each group is already known, there is a restriction on those deviations per group, which can be formulated by means of the underlying true values associated with a each group

$$\sum_{j=1}^{k} x_q^{g_q,j} = k \cdot \tilde{x}_q^{g_q,j} \quad \forall q, g \ , \tag{3}$$

or expressed in terms of the deviations within each group

$$\sum_{j=1}^{k} \Delta_q^{g_q,j} = 0 \quad \Longleftrightarrow \quad \Delta_q^{g_q,k} = -\sum_{j=1}^{k-1} \Delta_q^{g_q,j} \quad \forall q, g \ . \tag{4}$$

Depending on the employed microaggregation technique, further information on the regions in the data space in which the true underlying values are lying are deducible. Those regions are especially straightforward obtainable in case of Individual Ranking. Also in case of multivariate clustering one could identify such regions.

In the following two different approaches on obtaining meaningful estimators for the regression coefficients are presented.

3. A Nuisance Parameter Optimization Approach

This section discusses the optimistic estimation of the structural parameter when only considering favorable data situation. What is actually deemed as favorable is depending on the view taken.

A naive estimation approach just substitutes the true underlying data $\boldsymbol{x}$ with the microaggregated data $\tilde{\boldsymbol{x}}$, treating the microaggregated data as independent observations, resulting in the so-called *naive estimator* for $\boldsymbol{\beta}$. However, this estimator entirely neglects the nature of the data, it even explicitly rules out the imprecision. In the literature the properties of this naive estimator have been studied in the context of OLS estimation in order to improve it if necessary: Schmid and Schneeweiss (2008) have investigated the effects on the regression coefficient estimates in a general case when Individual Ranking was used, while in Schmid et al. (2007) Single-axis Sorting was considered with the response variable as the sorting variable. Further situations, including also external sorting variables, were discussed in Schmid (2007). Schmid and Schneeweiss (2005) conducted simulation studies, looking at the effect of different microaggregation techniques on the bias of the estimators. They prove that in some situations a bias correction is necessary and actually derive it for either the coefficient estimator or the error variance or even both. They also demonstrate that the corrected estimators are consistent and sometimes even the naive estimator has this property with respect to the true underlying data. However, they rely heavily on the fact that the OLS estimator is obtainable in a closed form, necessarily limiting their investigations to linear models.

3.1 Implications on the Score Function

When looking at the same task from the generalized linear regression perspective, a closed form is also obtainable for the classical linear regression as the estimation equations coincide, however the framework provides means for uniformly modeling different types of models for which neither

an OLS estimation is appropriate nor a closed form expression for the estimates is obtainable. Additionally, the fitting of the proposed corrections into the likelihood approach to obtain consistent estimators are tedious and not straightforwardly applicable; especially when thinking a step ahead when dealing with generalized linear regression. Yet again to obtain a concise estimate only one specific data constellation is to be deemed favorable, which is implicitly assessed by the likelihood.

In the following an extension is elaborated on stating this implicit assumption explicitly: One includes the underlying true values $\boldsymbol{x} \in \mathbb{X}$, or equivalently their deviation $\boldsymbol{\Delta}$ from the respective group mean, as nuisance parameters into the ideal likelihood and estimates them alongside:

$$\hat{\boldsymbol{\beta}} : \quad \ell(\boldsymbol{\beta}, \boldsymbol{x}; \boldsymbol{y}) \quad \longrightarrow \quad \max_{\boldsymbol{\beta}, \boldsymbol{x} \in \mathbb{X}}$$

The restrictions in (3) or (4) introduce a well perceived loss of freedom which in turn results in the fact that within each group there are $k-1$ nuisance parameters to estimate. For simplicity reason the restrictions as on the right in (4) are employed.

One should note that in the case of a classical linear model the resulting likelihood function to maximize is already a polynomial of grade 4 in the parameters, while taking the same approach in the logistic regression setting the function is not even polynomial.

As seen in (1) the score function consists of all observations' contributions. This overall summation may be re-ordered as one desires, and additionally even separately for the different score function parts; yet this comes at the cost that the score function is no longer directly expressible in the straightforward matrix notation as in (1). The general trick when looking at the score function is to exploit the grouping structure: For any $q = 1, \ldots, p$ the grouping with respect to the q^{th} covariate is employed when considering the score function part of β_q.

In the following the contribution $s(\beta_q)_g$ of such a group g to the score function with respect to β_q is given by including (4) into the ideal likelihood and then taking the respective derivatives:

$$\begin{aligned} s(\beta_q)_g = \frac{1}{\phi} \sum_{j=1}^{k} \left(\tilde{x}_q^{g_q,j} \left(y^{g_q,j} - \left(\beta_0 + \beta_q \tilde{x}_q^{g_q,j} + \sum_{\substack{r=1 \\ r \neq q}}^{p} \beta_r \left(\tilde{x}_r^{g_q,j} + \Delta_r^{g_q,j} \right) \right) \right) \right) \\ + \frac{1}{\phi} \sum_{j=1}^{k-1} \Delta_r^{g_q,j} \left[(y^{g_q,j} - y^{g_q,k}) - \beta_q \left(\Delta_q^{g_q,j} + \sum_{l=1}^{k-1} \Delta_q^{g_q,l} \right) \right. \\ \left. - \sum_{\substack{r=1 \\ r \neq q}}^{p} \beta_r (\tilde{x}_r^{g_q,j} + \Delta_r^{g_q,j} - \tilde{x}_r^{g_q,k} - \Delta_r^{g_q,k}) \right] . \end{aligned} \tag{5}$$

As each deviation is group specific for its respective covariate, the global score function $s(\Delta_q^{g_q,j})$ for the deviation $\Delta_q^{g_q,j}$ takes the following form:

$$\begin{aligned} s(\Delta_q^{g_q,j}) = \frac{\beta_q}{\phi} \left[(y^{g_q,j} - y^{g_q,k}) - \beta_q \left(\Delta_q^{g_q,j} + \sum_{l=1}^{k-1} \Delta_q^{g_q,l} \right) \right. \\ \left. - \sum_{\substack{r=1 \\ r \neq q}}^{p} \beta_r (\tilde{x}_r^{g_q,j} + \Delta_r^{g_q,j} - \tilde{x}_r^{g_q,k} - \Delta_r^{g_q,k}) \right] . \end{aligned} \tag{6}$$

As a necessary condition for the optimum the term in (6) needs to be zero. Under the assumption $\beta_q \neq 0$ it follows that the term in square brackets needs to equal zero. Furthermore, noting that the terms in square brackets are the same in (5) and (6), the score function $s(\beta_q)$ for β_q simplifies to:

$$s(\beta_q) = \frac{1}{\phi} \sum_{g=1}^{n/k} \sum_{j=1}^{k} \tilde{x}_q^{g_q,j} \left(y^{g_q,j} - \left(\beta_0 + \beta_q \tilde{x}_q^{g_q,j} + \sum_{\substack{r=1 \\ r \neq q}}^{p} \beta_r \tilde{x}_r^{g_q,j} \right) \right) - \frac{1}{\phi} \sum_{g=1}^{n/k} \sum_{j=1}^{k} \left(\tilde{x}_q^{g_q,j} \sum_{\substack{r=1 \\ r \neq q}}^{p} \beta_r \Delta_r^{g_q,j} \right) . \tag{7}$$

With Equation (7) it becomes obvious that, in the general case the estimate of β_q will be different to the one obtained by the naive estimator. However, when the grouping is the same for all covariates, the second line in (7) equals zero and the function is identical to the one obtained for the naive estimator. When looking at the respective equations for the intercept β_0 they are same in this approach and for the naive estimation. Nonetheless, as they are dependent on the other coefficients, the resulting estimates will still differ.

3.2 Notes on Consistency

When considering the limit behavior of the estimators in case of infinitely many observations, in particular their consistency, two situations are to be covered. In any case the number of observations prior to microaggregation increases, however the outcome will be dependent on whether the maximal group size is still fixed, i.e. k is independent of n, or if we increase k along with n, such that the ratio k/n is fixed. In the first case, as the number of observations increases and the group size does not change, the value of observations within a group converges to its group mean and therefore the deviations converge to zero. This then implies that the score estimation equation converges to the naive estimator equations, which converge to the ones obtain on the original data. Hence the nuisance parameters vanish in the converging point and the consistency of the one proposed in this section follows from the consistency of the general maximum likelihood estimator on the original data. However, such a property is in practice irrelevant, or even not desirable at all as it means that the privacy preserving effect of microaggregation vanishes in the limit. Hence the second case of a fixed ratio should be considered: In this case the observations will not converge towards the group means and equivalently the deviations will not vanish in the limit. This allows to conclude that the naive estimator is generally not a consistent one, as it neglects the non-vanishing deviations.

3.3 Maximax Optimization

The maximization of the likelihood as presented in the above subsection was formulated as an unconstrained optimization task with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Delta}$. Nonetheless, the task may also be formulated as a constrained optimization problem with respect to $\boldsymbol{\beta}$ and either the true underlying values $\boldsymbol{x}$ or the deviations $\boldsymbol{\Delta}$, with the microaggregation as constraint condition. As it appears more natural, in the following the choice of the restriction is switched to the ones on $\boldsymbol{x}$, as presented in (3). The previous formulation in terms of the $\boldsymbol{\Delta}$ already implements the microaggregation equality constraints into the target function by so-called *elimination*.

Furthermore, there might be additional restrictions on the original values which are specific for certain microaggregation techniques. For instance considering Individual Ranking: Here each variable is ordered and grouped individually and therefore additional bounds of the underlying true values for each variable are inferable as each underlying true value within a group must lie between the respective means of the neighboring groups. Also for MDAV, based on the Euclidean distance, one can add the heuristic restrictions that the true values are within a ball around each group mean with a radius of the minimal distance to any other group mean. By taking this optimization view, other – even external – constraints are implementable to restrict the task further.

In the here presented case the target function is still polynomial, while for generalized linear regression this does not necessarily hold. Therefore algorithms which can deal with such optimization tasks are required. As there already exist powerful algorithms to solve the maximization task in a generalized linear regression setting, instead of the direct maximization of the (log-)likelihood, the popular hill climbing strategy is employed. It consists of two major steps, which are to be repeated until a specified stopping criterion is reached:

1. Maximization of the likelihood for given $\hat{\boldsymbol{\beta}}^{(v-1)}$ estimates, such that all constraints on $\boldsymbol{x}$ are satisfied (not only microaggregation ones):

$$\hat{\boldsymbol{x}}^{(v)} = \arg\max_{\boldsymbol{x}} \ell(\boldsymbol{x}; \hat{\boldsymbol{\beta}}^{(v-1)}, \boldsymbol{y})$$

2. Maximization of the likelihood for given $\hat{\boldsymbol{x}}^{(v)}$ estimates:

$$\hat{\boldsymbol{\beta}}^{(v)} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \hat{\boldsymbol{x}}^{(v)}, \boldsymbol{y}) \ .$$

$\hat{\boldsymbol{x}}^{(v)}$ and $\hat{\boldsymbol{\beta}}^{(v)}$ denote the obtained estimates for $\boldsymbol{x}$ and $\boldsymbol{\beta}$ in the v-th iteration step. By this approach the complex optimization task is split into two different sub-tasks, which are in themselves easier to solve. The first is in general a non-convex optimization problem with linear constraints, while the second step is a simple estimation of a generalized linear regression for which standard software may be employed. One should note that the number of parameters to estimate is in general greater than the number of available observations which might result in more than a single solution. Furthermore, extensive care needs to be taken for choosing the initial values $\hat{\boldsymbol{\beta}}^{(0)}$: if they are set poorly the algorithm might converge only to a local maximum instead of the global one or might not converge at all, especially without the region constraints. Their introduction does make a difference to avoid some local maxima entirely, as it became visible in the simulation study in section 5.

4. A Partial Identification View

The previous sections rely implicitly on the assumption that the functions presented actually deserve their name as score function. As the number of nuisance parameters increases with the sample size, the situation is comparable to an example in Neyman and Scott (1948), therefore violating standard regularity conditions of maximum likelihood theory. They derive that in their situation the obtained estimators may not be consistent.

Instead of trying to correct the above functions and derive valid score functions, an ideological break is taken by looking at the task from a different angle, stressing the points of interest for generalized linear regression. The structural parameters $\boldsymbol{\beta}$ are of primary interest, while the underlying

values of the covariates are of minor interest or do not matter at all. In the spirit of partial identification only the available information is to be exploited, for instance the region constraints, while questionable assumptions are dropped, e.g. taking only the most favorable covariate constellation(s) into account in the model estimation. The partial identification approach in general leads to a set of estimates compatible with the available data. A natural approach is to calculate the so-called *collection regions*[6], or an outer approximation of it in form of a hypercube by calculating bounds on the coefficient estimates component wise. Those obtained may still be informative enough in practice. The collection region collects all such coefficients $\hat{\boldsymbol{\beta}}$ which are obtained as maximum likelihood estimates, or equivalently as zeros of the score function $s(\boldsymbol{\beta}; \boldsymbol{x}_0)$, for at least one feasible $\boldsymbol{x} \in \mathbb{X}$:

$$\hat{\boldsymbol{B}} := \left\{ \hat{\boldsymbol{\beta}} \mid \exists \boldsymbol{x}_0 \in \mathbb{X} : \ell(\hat{\boldsymbol{\beta}}; \boldsymbol{x}_0, \boldsymbol{y}) \geq \ell(\boldsymbol{\beta}; \boldsymbol{x}_0, \boldsymbol{y}) \, , \forall \boldsymbol{\beta} \in \mathbb{R}^{p+1} \right\} \overset{7}{=} \left\{ \hat{\boldsymbol{\beta}} \mid \exists \boldsymbol{x}_0 \in \mathbb{X} : s(\hat{\boldsymbol{\beta}}; \boldsymbol{x}_0) = 0 \right\} .$$

With this approach only the actually present information of the covariates is employed in the estimation of coefficients, it is guaranteed that the estimator of the unknown underlying micro data prior to microaggregation is contained within the resulting set, as well as the naive estimator. Furthermore, additional information on the covariates like marginal distributions are also includible.

When actually estimating the collection region $\hat{\boldsymbol{B}}$ or an appropriate approximation, an optimization perspective on the tasks proves fruitful once again. In order to obtain the component wise lower and upper bounds on $\hat{\boldsymbol{\beta}}$, the target function to minimize or maximize takes then a rather simple form:

$$\hat{\beta}_q \longrightarrow \min / \max \ .$$

Additionally to the constraints introduced by the microaggregation, namely the mean and region constraints, the score function constraint needs to be taken into account:

$$s_r(\hat{\boldsymbol{\beta}}; \boldsymbol{x}) = 0 \qquad \forall r \in \{0, \ldots, p\} \, ,$$

i.e. each score function part evaluated at the coefficient vector $\hat{\boldsymbol{\beta}}$ and a feasible $\boldsymbol{x}$.

Please note that in contrast to the previous section, the $\boldsymbol{x}$ are only indirectly subject to the optimization. They are allowed to vary freely within their respective bounds, as long as the summation restrictions induced by the microaggregation and any other constraints on them are satisfied. There is no further plausibility assessment on their actual values, as it was employed when optimizing them in the light of the log-likelihood concomitantly or when constructing corrected estimators.

As in general the constraint on the score function has a not negligible complexity, especially as it is not linear in the parameters, the equality constraint may be better incorporated into the target function in terms of a penalty. This leads to the following target function:

$$\hat{\beta}_q \pm \sum_{r=0}^{p} \lambda_r (s_r(\hat{\boldsymbol{\beta}}; \boldsymbol{x}))^2 \ \longrightarrow \ \min / \max \, ,$$

where the sign before the sum is chosen appropriately[8] and λ_r are the so-called *penalty parameters*. If λ_r increases each deviation of the evaluated score function from zero is penalized to a greater extent. Therefore by sufficiently large enough λ_r the deviation is numerically forced onto zero.

Another benefit of those views is the ability to check for any given vector $\boldsymbol{\beta}^*$ if it is included in the feasible region $\hat{\boldsymbol{B}}$ of regression coefficients, which is an optimization task in $\boldsymbol{x}$ only.

6. cf. Schollmeyer and Augustin (2015, sec. 3.2) for an overview of other types of identification regions.
7. This equivalence holds for generalized linear regression as the log-likelihood is concave; cf. Wedderburn (1976).
8. The addition is chosen when minimizing and subtraction when maximizing the component.

n	Average RMSE			
	Truth	IR	SaS	MDAV
100	0.453	0.667 (0.459)	0.797 (0.797)	0.461 (0.462)
250	0.268	0.352 (0.269)	0.469 (0.469)	0.270 (0.270)

Table 1: Average root mean squared error (RMSE) for the maximax-like approach in different microaggregation settings; values in parentheses are the respective value for the naive estimator

5. Simulation Results

Besides the theoretical investigations, a simulation study is conducted with the aim to visualize the effects of microaggregation and display the adequacy of the proposed methods.[9] The simulation was carried out with the statistical software *R* (R Core Team, 2016)[10].

The classical linear model in the generalized linear regression setting is considered, for simplicity reasons with two independent covariates. Exemplary, the microaggregation techniques of Single-axis Sorting (SaS), Individual Ranking (IR) and MDAV are employed onto the setting of two independent covariates; the dependent variable is left un-aggregated. The regression coefficients are estimated by means of the hill climbing algorithm with all available constraints in the sense of section 3 and for the component wise bounds by taking the partial identification view as presented in section 4. The covariates are each drawn from a uniform distribution on $[0, 10]$, while the response variable is obtained by inducing a dependency structure based on $\boldsymbol{\beta} = (1, 0.5, 1.9)^T$ and adding white noise with variance $\sigma^2 = 4$. To study the behavior when more observations are available, the number of observations varied between $n = 100$ and 250, while the aggregation size was kept fixed with $k = 5$. Each setting was repeated 1000 times.

Under consideration were the different proposed methods: optimization with only microaggregation equality constraints, optimization with microaggregation equality constraints and region inequality constraints and finally component wise coefficient min/max estimation of the collection region. However, in a preliminary run, it turned out that the first method, which ignores de-facto available information, is also numerically unstable as the results are very highly dependent on the choice of the initial value, which is a well known difficulty when applying a hill climbing strategy.[11] However as soon as the additional region constraints entered the task, this was no longer the case and the obtained solution was a reliable one. Therefore in the here presented simulation the optimization with just the microaggregation equality constraints is left out.

In Table 1 some results of the simulation are summarized: For the concise estimates their average root mean squared error (RMSE) is reported. As can be seen for the average RMSE, in the case of Individual Ranking the maximax-like approach performs poorer in comparison to the naive estimator. For Single-axis Sorting the estimator coincide, as theoretically shown in section 3, which

9. The code and supplementary files are available on request.
10. The employed optimizer *SLSQP* (Kraft, 1994) is provided by the R package NLoptr (Johnson, 2014).
11. It was found that most of the times in cases when the initial values for $\boldsymbol{\beta}^{(0)}$ were considerably far away from the true ones, a solution was returned, which was easily improvable.

practically also holds for the MDAV approach[12]. Furthermore, with higher n the average RMSE is smaller, reflecting the considerations in section 3.2. When comparing the average RMSE obtained on the aggregated data to the one on the original, one finds that Single-axis Sorting produced the highest discrepancy, while the others yield quite comparable results. This is mainly due to the fact that Singe-axis Sorting destroys any multivariate structure within the data. In general it is found that the maximax-like approach is too optimistic resulting in poorer performance, and that the naive estimator should be preferred if the user insists on a guaranteed precise solution. The simulation essentially confirms the analytical results of section 3 and the corresponding ones in Schmid and Schneeweiss (2005).

For the outer approximation of the identification region the volume of the coefficient box was calculated. For those the results are also in line with the findings for average RMSE, as in case of Single-Axis Sorting the boxes are considerably large ($n = 100$: 38.031; $n = 250$: 1.530), while for Individual Ranking and MDAV the estimators are numerically point identified. Nonetheless, for Single-Axis Sorting the box volume shrinks as n increases. For any methods in any repetition the naive estimator is indeed included within the box or coincides with the point identified estimators.[13] The tight boxes are mainly due to the exploitation of the guessable region constraints for Individual Ranking and MDAV.

6. Concluding Remarks

In the light of protecting sensitive data, especially micro data, and availability of suitable methods for protection, the usability of such protected data should no longer be neglected. In this paper a general investigation of the effect of different microaggregation methods on the outcome of regression coefficient estimation in generalized linear regression was started. For generic microaggregation it was demonstrated how the ideal likelihood as the core of generalized linear models can be used as basis in estimation: on the one hand by introducing nuisance parameters concerning the true underlying values and on the other hand by a partial identification view resulting in a set of reachable values. In the present paper for the partial identification approach an outer approximation of the actual set was given, yet it may be further refined. Furthermore, the basic ideas were demonstrated on a classical linear regression, expressed in the framework of generalized linear regression. However, the developed concepts are general and therefore may be employed to other models, for instance the logistic regression.

Acknowledgments

We are grateful for the remarks of three anonymous reviewers, also stimulating further research.

References

J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002. doi:10.1109/69.979982.

12. The minor difference for MDAV is subject to the heuristic region constraint used.
13. Interestingly the estimator on the original data was not always contained within the calculated box, which might be subject to the employed optimizer. This clearly needs further investigations as theory guarantees the inclusion.

Eurostat and European Statistical System. The European Statistics Code of Practice, 2011. doi:10.2785/18474.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications*. Springer, Berlin, 2013.

S. G. Johnson. *The NLopt Nonlinear-optimization Package*, 2014. URL http://ab-initio.mit.edu/nlopt.

D. Kraft. Algorithm 733: TOMP–Fortran modules for optimal control calculations. *ACM Transactions on Mathematical Software*, 20(3):262–281, 1994. doi:10.1145/192115.192124.

C. F. Manski. *Partial Identification of Probability Distributions*. Springer, Berlin, 2003.

J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948. doi:10.2307/1914288.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org.

M. Rosemann, D. Vorgrimler, and R. Lenz. Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten. *Allgemeines Statistisches Archiv*, 88(1):73–99, 2004. doi:10.1007/s101820400160.

M. Schmid. *Estimation of a Linear Regression with Microaggreated Data*. Verlag Dr. Hut, Munich, 2007.

M. Schmid and H. Schneeweiss. The effect of microaggregation procedures on the estimation of linear models: A simulation study. Technical Report 443, Institut für Statistik, Sonderforschungsbereich 386, München, 2005. URL https://epub.ub.uni-muenchen.de/1831.

M. Schmid and H. Schneeweiss. Estimation of a linear model in transformed variables under microaggregation by individual ranking. *AStA Advances in Statistical Analysis*, 92(4):359–374, 2008. doi:10.1007/s10182-008-0087-9.

M. Schmid, H. Schneeweiss, and H. Küchenhoff. Estimation of a linear regression under microaggregation with the response variable as a sorting variable. *Statistica Neerlandica*, 61(4):407–431, 2007. doi:10.1111/j.1467-9574.2007.00366.x.

G. Schollmeyer and T. Augustin. Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *International Journal of Approximate Reasoning*, 56:224–248, 2015. doi:10.1016/j.ijar.2014.07.003.

L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002. doi:10.1142/S0218488502001648.

R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1):22–32, 1976. doi:10.2307/2335080.

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