(Generalized) Linear Regression on Microaggregated Data

Paul Fink, **Thomas Augustin**

Department of Statistics

LMU Munich

July 11th, 2017

ISIPTA '17 / ECSQARU 2017 Lugano, Switzerland

Biography

Paul Fink	PhD student in group *Foundations of Statistics and Their Applications* M. Sc. in Statistics Strong interest in the analysis of so-called deficient data, e.g. anonymized data
Thomas Augustin	Head of group *Foundations of Statistics and Their Applications*

Anonymization: Background

- **General Aim**: Sharing of micro data to a broader audience, e.g. in Official Statistics
- **Issue**: Protection of sensitive information to prohibit disclosure of records ($\longrightarrow$ privacy)
- **Solution**: Anonymization in a way that balance
  1. the privacy requirement and
  2. the contained statistical quality
- Microaggregation as a set of methods for anonymization of metrical variables

How severe does the anonymization affect the analysis outcome?

Microaggregation

Typical structure of microaggregation techniques

Grouping: Partition individual records of the micro data into clusters such that records within a cluster are similar and each cluster contains at least $k \geq 3$ records

Aggregation: Replacement of each individual record within a cluster by the cluster's characteristic value, e.g. mean or median

Many microaggregation techniques available, differing mostly in grouping step

Representation as data transformation:

$$\boldsymbol{x} \xrightarrow{m} \tilde{\boldsymbol{x}}$$

Microaggregation – Example ($k = 3$)

Original data $\boldsymbol{x}$

ID	Turnover	Profit	...
1	70.951	4.270	
2	15.610	−3.029	⋮
3	105.593	−4.160	
4	80.929	−2.215	
5	17.156	−9.941	⋮
6	6.020	2.140	
7	102.936	−13.475	
8	49.407	−6.167	
9	143.424	−6.826	⋮
10	59.793	9.404	

Microaggregation – Example ($k = 3$)

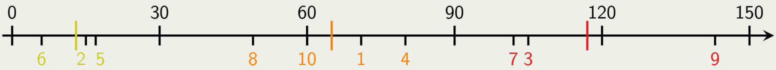
Original data $\boldsymbol{x}$

ID	Turnover	Profit	…
1	70.951	4.270	
2	15.610	−3.029	⋮
3	105.593	−4.160	
4	80.929	−2.215	
5	17.156	−9.941	
6	6.020	2.140	⋮
7	102.936	−13.475	
8	49.407	−6.167	
9	143.424	−6.826	⋮
10	59.793	9.404	

Individual Ranking $\tilde{\boldsymbol{x}} = m(\boldsymbol{x})$

ID	Turnover	Profit	…
1	65.270	5.271	
2	12.929	−3.893	⋮
3	117.318	−3.893	
4	65.270	−3.893	
5	12.929	−10.081	
6	12.929	5.271	⋮
7	117.318	−10.081	
8	65.270	−3.893	
9	117.318	−10.081	⋮
10	65.270	5.271	

Turnover:



'Inverse' Microaggregation – Example ($k = 3$)

Anonymized data $\tilde{\boldsymbol{x}}$

ID	Turnover	Profit	...
1	65.270	5.271	
2	12.929	−3.893	⋮
3	117.318	−3.893	
4	65.270	−3.893	
5	12.929	−10.081	⋮
6	12.929	5.271	
7	117.318	−10.081	
8	65.270	−3.893	
9	117.318	−10.081	⋮
10	65.270	5.271	

'Inverse' Microaggregation – Example ($k = 3$)

Anonymized data $\tilde{\boldsymbol{x}}$

ID	Turnover	Profit	...
1	65.270	5.271	
2	12.929	−3.893	⋮
3	117.318	−3.893	
4	65.270	−3.893	
5	12.929	−10.081	⋮
6	12.929	5.271	
7	117.318	−10.081	
8	65.270	−3.893	
9	117.318	−10.081	⋮
10	65.270	5.271	

Compatible data $\boldsymbol{x}_1$: $m(\boldsymbol{x}_1) = \tilde{\boldsymbol{x}}$

ID	Turnover	Profit	...
1	73.316	9.039	
2	15.214	−4.874	⋮
3	164.674	−2.066	
4	47.416	−6.369	
5	7.849	−13.106	⋮
6	15.724	3.691	
7	103.918	−6.923	
8	75.067	−2.263	
9	83.362	−10.214	⋮
10	65.281	3.083	

'Inverse' Microaggregation – Example ($k = 3$)

Anonymized data $\tilde{\boldsymbol{x}}$

ID	Turnover	Profit	...
1	65.270	5.271	
2	12.929	−3.893	⋮
3	117.318	−3.893	
4	65.270	−3.893	
5	12.929	−10.081	⋮
6	12.929	5.271	
7	117.318	−10.081	
8	65.270	−3.893	
9	117.318	−10.081	⋮
10	65.270	5.271	

Compatible data $\boldsymbol{x}_2$: $m(\boldsymbol{x}_2) = \tilde{\boldsymbol{x}}$

ID	Turnover	Profit	...
1	53.567	4.247	
2	10.763	−8.688	⋮
3	109.089	−9.058	
4	69.812	−1.507	
5	13.955	−9.480	⋮
6	14.069	6.509	
7	133.563	−9.999	
8	79.483	3.681	
9	109.302	−10.764	⋮
10	58.218	5.057	

'Inverse' Microaggregation – Example ($k = 3$)

Anonymized data $\tilde{\boldsymbol{x}}$

ID	Turnover	Profit	...
1	65.270	5.271	
2	12.929	−3.893	⋮
3	117.318	−3.893	
4	65.270	−3.893	
5	12.929	−10.081	⋮
6	12.929	5.271	
7	117.318	−10.081	
8	65.270	−3.893	
9	117.318	−10.081	⋮
10	65.270	5.271	

Compatible data $\boldsymbol{x}_2$: $m(\boldsymbol{x}_2) = \tilde{\boldsymbol{x}}$

ID	Turnover	Profit	...
1	53.567	4.247	
2	10.763	−8.688	⋮
3	109.089	−9.058	
4	69.812	−1.507	
5	13.955	−9.480	⋮
6	14.069	6.509	
7	133.563	−9.999	
8	79.483	3.681	
9	109.302	−10.764	⋮
10	58.218	5.057	

Microaggregated data induce set of compatible data:

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

(Generalized) Linear Regression

Modeling the conditional expectation $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})$ by a (transformed) linear predictor $\boldsymbol{x}\beta$.

Estimation of the parameter of interest β by maximum likelihood:

Log-likelihood: $\ell(\boldsymbol{\beta}; \boldsymbol{x}, \boldsymbol{y}) \quad \longrightarrow \quad \max_{\boldsymbol{\beta}}$

$$\Updownarrow$$

Score function: $s(\beta; \boldsymbol{x}, \boldsymbol{y}) = \dfrac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} = 0$

(Generalized) Linear Regression on Microaggregated Data

Analysis of contained statistical quality with respect to (generalized) linear regression

for microaggregated covariate(s) $\tilde{\boldsymbol{x}}$

on a non-microaggregated response $\boldsymbol{y}$.

Of interest is the connection between $\boldsymbol{y}$ and the unobserved $\boldsymbol{x}$!

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

→ Nuisance Parameter Optimization

→ Partial Identification

Nuisance Parameter Optimization

Treating of the underlying true values as nuisance parameters

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

In linear regression the *nice* score function structure reduces the complexity of the optimization task.

Incorporating additional (in)equalities specific for the applied microaggregation technique $\longrightarrow$ More concise estimates

Partial Identification

Aim: Estimating the collection region

$$\hat{\boldsymbol{B}} := \{\hat{\beta} \mid \exists \boldsymbol{x}_0 \in \mathbb{X} : s(\hat{\beta}; \boldsymbol{x}_0, \boldsymbol{y}) = 0\}$$

Estimation of component wise lower and upper bounds on $\boldsymbol{\beta}$:

$$\hat{\beta}_{\boldsymbol{q}} \longrightarrow \min / \max$$

such that

- all score functions requirements and
- additional (in)equalities specific for the applied microaggregation technique

are satisfied.

Solving via penalized optimization approach:

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

Summary and Outlook

- Microaggregated data induce set of compatible data

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

→ Nuisance Parameter Optimization

→ Partial Identification

- Simulation study with three microaggregation techniques

- Analysis of contained statistical quality with respect to generalized linear regression, e.g. logistic regression
- Analysis on the influence of the microaggregation technique