

Guest Editorial “Econometrics of Anonymized Micro Data”

Individual data which are collected by the statistical office or other governmental institutions offer a rich source of information which could be used for applied research, especially in economic and social sciences. However these data cannot be released freely due to confidentiality reasons. This is especially true if households or firms are obliged by law to provide the information which is the case, for example, for Germany. And in particular data from firms may contain information on, say, profit, sales or R&D which should not be distributed in public. In many countries researchers have been given the opportunity to use these micro data which are “formally” anonymized, that is, where any identifier such as name or address of household or firm has been removed from the data set. However, individual firms or households could still be re-identified if they show unusual characteristics: a household with high income and more than seven children or a firm with maximum sale within the automobile industry. To avoid such re-identification statistical offices have used alternative strategies: (i) Researchers have to sign a contract which implies severe punishment in case of any effort to re-identify individual firms or persons. (ii) Researchers have to do all their calculations under control of the statistical office. (iii) Data could be provided as “Scientific Use Files” which means that the micro data have been processed in such a way that re-identification of individual units may still be possible but would involve high cost. Moreover these data are given only to researchers who guarantee that they use these data in their own scientific work.

For the third alternative data have to be anonymized by statistical procedures such as rank swapping, micro aggregation or addition of noise in case of quantitative variables, and by post-randomization in case of qualitative variables. There has been a large body of publications discussing adequate approaches of masking the data. If such procedures guarantee satisfactory protection of data, the question naturally arises whether these anonymized data can be used instead of the original data and how reliable estimates from this alternative data set would be. The German Statistical Office has initiated a project which analyses the possibility of producing scientific use files which satisfy the conditions of “factual anonymization” as laid down by German law.¹ As far as we can see this project is the first systematic attempt not only to analyze protection against re-identification but also to rate the quality of such data with regard to statistical analyses. Since the project in particular considers economic data, the main emphasis was laid on the estimation of microeconomic models. However these models, in particular the linear model and the probit model, are used frequently in other areas of applied research, too.

First versions of most contributions collected in this issue under the title “Econometrics of Anonymized Micro Data” have been presented at the conference “Econometric Analysis of Anonymised Firm Data” organized by The Institute for Applied Economic Research, Tübingen, on 18–19 March 2004 and which is related to the above mentioned project. All papers are mainly methodological. Some however also use empirical examples to illustrate

¹ The project “Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten” has been financially supported by the German Ministry of Research and Education. “Factual anonymization” is defined in § 16 (6) of the German “Bundesstatistik-Gesetz”. Scientific use files are obtainable from the Research Data Center of the German Statistical Office.
See <http://www.forschungsdatenzentren.de/nutzungsantrag.asp>

the results. One paper replicates a former empirical study using now anonymized data and compares the outcomes with those originally obtained. For anonymization only the following procedures are considered: noise addition (and some variant termed resampling), microaggregation and post-randomization. These approaches have been found particularly useful in both protecting the data and leaving enough informational content for reliable estimation of the stochastic models.

We close this editorial by shortly introducing the five papers: **Lechner and Pohlmeier** consider linear and nonlinear models in which the explanatory variables have been masked by noise addition which is formally equivalent to the problem of “errors in variables”. The authors show that the simulation-extrapolation (SIMEX) estimator is a convenient tool for consistent estimation of parametric and nonparametric model specifications. **Schmid and Schneeweiss** analyse the effect of microaggregation procedures on the estimation of the linear model. Their results indicate that microaggregation when related to the dependent variable may be a sound procedure for some variants of this procedure whereas others imply an asymptotic bias for the coefficient estimator. **Ronning, Rosemann and Strotmann** consider the probit model for the case that the observed binary dependent variable has been anonymized by post-randomization. They show that consistent estimation is possible if the estimator is adequately modified. **Gottschalk** uses a nonparametric kernel technique combined with resampling for the anonymization of the data. These data are then used for estimation of linear and non-linear models. The author shows that an adaptive procedure will give most satisfactory results. **Wagner** re-estimates two of his own empirical studies which were both based on the “Hannover Firm Panel Study”. He considers the effect of (partial) microaggregation on results regarding both estimation and testing employing a battery of microeconomic models.

Winfried Pohlmeier (Universität Konstanz)

Gerd Ronning (Universität Tübingen)

Joachim Wagner (Universität Lüneburg)