First publ. in: Advances in Statistical Analysis 92 (2008), 4, pp. 351-357

Microeconometrics and disclosure control

Winfried Pohlmeier · Gerd Ronning

1 Introduction

Most micro data are collected by national statistical offices and, in turn, require confidential treatment so that published data tables are usually restricted to aggregate figures (e.g., household income or industrial production). Such data are, of course, also available in disaggregated form: For example, income may be differentiated by region or occupation, and production may be available for a single industry or for a single region. Additionally, tables of two or more dimensions are often provided, which, for example, may display income by region and occupation or production by industry and region. Certain cells in these tables, however, cannot be revealed due to the intrinsic possibility of identifying individuals or, more frequently, specific firms. A famous example is automobile production within a certain region; all production is typically due to one large enterprise, and the corresponding publication of the production of local automobile manufactures will clearly compromise this large firm's confidentiality.

Of course, modern empirical research in the social sciences is not content with statistical information provided by tables but asks for the individual ("micro-") data.

The editors like thank the referees of the papers collected in this volume for a thorough and swiftly refereeing. Without their support we would not have been able to publish the results of the conference "Methodical Aspects of the Anonymization of Panel Data", without significant delay.

W. Pohlmeier (✉)
Department of Economics and Center for Quantitative Methods and Survey Research (CMS), University of Konstanz, 78457 Konstanz, Germany
e-mail: winfried.pohlmeier@uni-konstanz.de

G. Ronning
Wirtschaftswissenschaftliche Fakultät, Universität Tübingen, Mohlstrasse 36, 72074 Tübingen, Germany
e-mail: gerd.ronning@uni-tuebingen.de

In particular, econometricians have developed a powerful methodological tool box under the heading "microeconometrics". Nowadays not only the interest of empirical researchers but also the increasing public awareness that research based policy advice requires adequate data sources has launched the increasing demand for micro data.

The standards governing the use of micro data obtained from national statistical offices is organized quite differently in different countries. In some countries micro datasets are given to researchers without any prior manipulation or masking but with the stipulation that a huge fine is to be paid in the event that the data is turned to ill-use. This is the case, for example, in Switzerland.[1] In most countries, however, researchers are required to send their computer programs to the statistical office ("remote access") or could even be required to be physically present within the statistical office in question, to prevent the identification of any single observational unit.

This volume deals with a third alternative: data are made available after having been manipulated so that the risk of re-identification is negligible and/or the cost of disclosure is prohibitive. Depending on the extent of the disclosure limitation applied, these data files are called "public use files" or "scientific use files", the latter containing more informational content but which also implies a higher risk of re-identification.

For a long time only disclosure control has been of primary interest when providing micro data,[2] whereas the degree of usefulness of such data for statistical research was put aside. Of course, the higher the extent of protection applied to the data, the less useful that data will be for empirical researchers. For example, swapping and rank swapping have historically been considered to be especially useful for data protection, but recent studies show that estimation results based on such anonymized data will be severely biased.[3]

However, there has also been some effort to recognize the effects of anonymization in estimating stochastic models. A well-known contribution by Kim (1986) proposed that noise be added in such a way that first and second moments of the data set remain unchanged. This implies that estimation of linear models will be not affected by the manipulation of data. Lechner and Pohlmeier (2003) were among the first to draw attention to the fact that addition of noise will result in the famous errors-in-variables model and should be estimated accordingly using, for example, instrumental variables. In the same paper, they also considered estimation of linear models when data were anonymized by micro-aggregation which lends itself to generalized least squares estimation.

The latter mentioned paper was written as part of a German project which studied the possibilities of providing researchers with anonymized micro data in the form of scientific use files.[4] Special emphasis in the two projects was given to the ques-

[1] Personal communication with Michael Lechner, University of St. Gallen.

[2] See, for example, Willenborg and de Waal (2001) for an overview and the program of the most recent conference regarding the topic: Privacy in Statistical Databases 2008, Istanbul, September 2008.

[3] See, e.g., Ronning et al. (2005), p. 65.

[4] The two projects "Faktische Anonymsierung wirtschaftsstatistischer Einzeldaten" (2002–2005) and "Wirtschaftsstatistische Paneldaten und faktische Anonymisierung" (2006–2008) were financed by the Bundesministerium für Bildung und Forschung. The first project considered only cross-sectional data, and the second project extended research to the case of panel data.

tion of how estimation from anonymized data can be organized so that estimates remain unbiased or at least consistent while sacrificing a minimal amount of efficiency. A first conference on this topic was organized in March 2004 at the Institute of Applied Economic Research (IAW), Tübingen, under the title "Econometric Analysis of Anonymised Firm Data". Results were edited by Pohlmeier et al. (2005) in a special issue of "Jahrbücher für Nationalökonomie und Statistik/Journal of Economics and Statistics". In 2007 a follow-up conference titled "Methodical Aspects of the Anonymization of Panel Data" took place at the IAW in Tübingen, where preliminary versions of the six papers collected in this special issue were presented.

2 Anonymization techniques considered in this volume

All papers consider one of the following anonymization procedures:[5]

- Micro-aggregation
- Noise Addition
- Data Blanking
- Multiple Imputation

These are briefly detailed in the following.

Micro-aggregation

Micro-aggregation first forms a (small) group of observational units and then replaces the corresponding original values with the group average. If this procedure is applied jointly to a set of variables, it is very similar to (one-step) cluster analysis. However, an optimal solution regarding grouping is very complex.[6] Moreover, joint micro-aggregation does alter the correlation structure significantly. These issues led to the discussion of micro-aggregation separately for each variable, which is termed "micro-aggregation by individual ranking". Intuitively, this approach will modify values for most units only slightly, but, of course, the risk of re-identification is then correspondingly higher. The paper by Schmid and Schneeweiß considers this approach.

Noise addition

Addition of noise also can be applied jointly to all variables or to each variable separately. Until recently, anonymization using additive noise has been of primary concern. The above mentioned projects, however, very clearly showed that multiplicative noise is much more efficient in protecting both small and large firms: If sales of two firms amount to 100,000 Euros for A and 10 million Euros for B, an additive noise of, say, 5,000 Euros will modify sales of A considerably, whereas those of B will remain almost unchanged. More formally, let X be the variable indicating the original value,

[5] See, e.g., Ronning et al. (2005) for an extensive discussion of the various anonymization procedures.

[6] See, for example, Oganian and Domingo-Ferrer (2001).

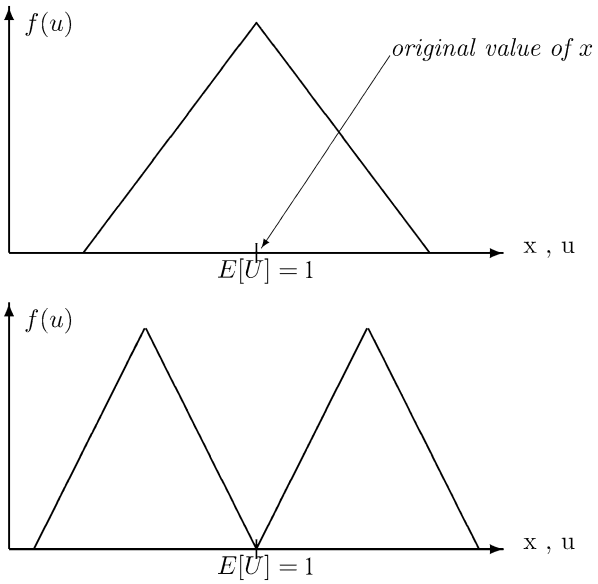

Fig. 1 Multiplicative noise using mixture distribution

X^a the anonymized variable, and U the noise variable. Then additive noise is defined by

$$X^a = X + U \quad \text{with } E[U] = 0,$$

and multiplicative noise is given by

$$X^a = X \cdot U \quad \text{with } E[U] = 1.$$

In the multiplicative case both sales will be altered by a certain percentage so that both small and large values will be modified in a similar way. Most importantly, large firms will be much better protected.

Protection by multiplicative noise may be further improved by using a bimodal mixture distribution instead of the usual unimodal distribution.[7] As Fig. 1 illustrates, in the case of the mixture distribution, most anonymized values will be further away from the original values than as would be the case if a unimodal distribution was used. We note in passing that this bimodal mixture distribution may also be characterized by

$$X^a = X \cdot (1 + \delta D + \varepsilon),$$

where the binary random variable D satisfies

$$D = \begin{cases} +1 & \text{with probability } \alpha, \\ -1 & \text{with probability } 1 - \alpha, \end{cases}$$

[7] Use of mixture distributions in data protection was first proposed by Roque (2000). A simplified procedure was later described by Yancey et al. (2002).

and ε is white noise. The parameter δ describes the percentage change, and the mixture parameter α normally will be set to 0.5 in anonymization.

Four papers in this volume consider addition of noise techniques. See Biewen, Nolte, and Rosemann, Ronning and Rosemann, Flossmann, and Nolte, Biewen, and Ronning.

Data blanking

Blanking is the easiest method of disclosure limitation. Individual observations which could subject entities to a significant risk of re-identification are simply erased from the data set. This can be done partly, so that only the variables at risk are blanked—or it can be carried out on all variables. If a dependent variable is subject to data blanking, estimation takes the form of the well-known sample selection problem, where the selection rule is given by the legal requirements. The contribution by Flossmann and Nolte considers econometric options for data blanking in combination with multiplicative noise.

Multiple imputation

Multiple Imputation was originally proposed for statistical analysis of data with missing values by Rubin (1976, 1987). Rubin (1993) then suggested the use of this approach as a device for data protection. The idea is to substitute all observations with "missing values" which are completely synthetic and therefore satisfy the objective of disclosure control in a perfect manner. For applications of this approach, see, e.g., Raghunatan et al. (2003). The paper by Drechsler, Dundler, Bender, Rässler, and Zwick is devoted to this approach.

3 A short characterization of contributions to this volume

The paper by Schmid and Schneeweiß considers the problem that researchers often transform variables when specifying their models. For example, age squared is often used in labor econometrics. The question then arises whether the anonymization should be done for the transformed variable (which would make anonymization effort larger) or whether the already anonymized variable can be transformed and then used in the estimation procedure. The authors show that in the special case of separate micro-aggregation ("individual ranking") estimation of the linear model is not affected, at least asymptotically, if the transformed anonymized variables are used.

Biewen, Nolte, and Rosemann compare two different approaches of *multiplicative* noise. They apply the Simulation Extrapolation method (SIMEX) which was originally proposed by Cook and Stefanski (1994) in order to correct for bias due to *additive* measurement error (See also Carroll et al. 2006). In the first approach they reformulate the multiplicative noise model as an additive one and use the SIMEX estimator for the standard additive case. In the second approach they use the model with untransformed multiplicative noise. The simulation study illustrates their results for a Probit model.

Ronning and Rosemann consider the case of *additive* errors. They also apply the SIMEX procedure assuming that in a simple linear regression model the errors for the dependent and the independent variables are correlated. They modify the SIMEX procedure for this situation and show that it would also work in case of many regressors. Moreover they demonstrate in a simulation study that neglecting correlation will lead to estimates which may be worse than those from the naive estimator which completely disregards measurement errors.

Flossmann and Nolte combine two separate disclosure limitation techniques, blanking and multiplication of independent noise, in order to better protect the original dataset. They combine the multiplicative Simulation-Extrapolation (M-SIMEX) approach with both the Inverse Probability Weighting (IPW) approach (going back to Horwitz and Thompson 1952) and with matching methods. They show that noise multiplication combined with blanking as a masking procedure does not necessarily lead to a severe reduction in the estimation quality when using these estimation procedures.

Biewen and Ronning also consider the special case of a mixture distribution of *multiplicative* noise used for the anonymization of panel data. They consider bias of the naive least-squares estimator ("within estimator") and derive correction formulae in order to obtain consistent estimates. They show that in short panels these formulae would have to take into account the possible autocorrelation of regressors which will typically be the case in economic panel data. However, this implies that the (heterogeneous) autocorrelation for different firms would have to be estimated. Alternatively, the SIMEX approach is tried which also does not work satisfactorily in the presence of autocorrelation.

Finally Drechsler et al. describe how multiple imputation can be used for the protection of micro data. They present an application of Rubin's (1993) idea of generating synthetic datasets from existing confidential survey data for public release. A set of variables from the 1997 wave of the German IAB Establishment Panel is used to evaluate the quality of the approach. The comparison of results for original data and anonymized data shows that valid inferences can be obtained using the synthetic datasets in this context, while guaranteeing confidentiality for the survey participants.

References

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement Error in Nonlinear Models. A Modern Perspective, 2nd edn. Chapman and Hall, London (2006)

Cook, J.R., Stefanski, L.A.: Simulation-extrapolation estimation in parametric measurement error models. J. Am. Stat. Assoc. **89**, 1314–1328 (1994)

Horwitz, D., Thompson, D.: A generalization of sampling without replacement from a finite population. J. Am. Stat. Assoc. **47**, 663–685 (1952)

Kim, J.J.: A method for limiting disclosure in microdata based on random noise and transformation. In: Proceedings of the Section on Survey Research Methods, pp. 303–308. American Statistical Association, Alexandria (1986)

Lechner, S., Pohlmeier, W.: Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten (Estimation of econometric models on the basis of anonymized data). In: Forum der Bundesstatistik, vol. 42, pp. 115–137. Statistisches Bundesamt, Wiesbaden (2003)

Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control. In: Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, 14–16 March 2001

Pohlmeier, W., Ronning, G., Wagner, J.: Econometrics of anonymized micro data. Jahrb. Nationalökonomie Stat. J. Econ. Stat. **225**(5) (2005)
Raghunatan, T., Reiter, J., Rubin, D.: Multiple imputation for statistical disclosure limitation. J. Stat. **19**, 1–16 (2003)
Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M., Vorgrimler, D.: Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. Statistik und Wissenschaft, vol. 4. Statistisches Bundesamt, Wiesbaden (2005)
Roque, G.M.: Masking microdata files with mixtures of multivariate normal distributions. PhD Dissertation, University of California-Riverside (2000)
Rubin, D.B.: Inference and missing data. Biometrika **63**, 581–592 (1976)
Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, New York (1987)
Rubin, D., Discussion. Statistical disclosure limitation. J. Off. Stat. **9**(2), 461–468 (1993)
Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Springer Lecture Notes in Statistics, vol. 155. Springer, Berlin (2001)
Yancey, W.E., Winkler, W.E., Creecy, R.H.: Disclosure risk assessment in perturbative microdata protection. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases, pp. 135–152. Springer, New York (2002)