Econometrics of Anonymized Micro Data

Edited by

Winfried Pohlmeier, Gerd Ronning, Joachim Wagner

With Contributions by

Sandra Gottschalk, Mannheim
Sandra Lechner, Konstanz
Winfried Pohlmeier, Konstanz
Gerd Ronning, Tuebingen
Martin Rosemann, Tuebingen
Matthias Schmid, Munich
Hans Schneeweiss, Munich
Harald Strotmann, Tuebingen
Joachim Wagner, Lueneburg

Lucius & Lucius Verlagsgesellschaft Stuttgart 2005

Anschriften der Herausgeber des Themenheftes

Dr. Winfried Pohlmeier
Department of Economics
Box D124
University of Konstanz
D-78457 Konstanz (Germany)

E-mail: winfried.pohlmeier@uni-konstanz.de

Professor Dr. Gerd Ronning
Wirtschaftswissenschaftliche Fakultaet
Universitaet Tuebingen
Mohlstr. 36 (Room 511)
D-72074 Tuebingen (Germany)

E-mail: gerd.ronning@uni-tuebingen.de

Professor Dr. Joachim Wagner
University of Lueneburg
Campus 4.210
D-21332 Lueneburg (Germany)

E-mail: wagner@uni-lueneburg.de

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar

ISBN 3-8282-0325-6

Satz: Mitterweger & Partner Kommunikationsgesellschaft mbH, Plankstadt
Druck und Bindung: Neumann Druck, Heidelberg
Printed in Germany

Inhalt / Contents

Abhandlungen/Original Papers

Buchbesprechung/Book Review

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 guarantuee 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 microeconometric 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 microeconometric models.

Winfried Pohlmeier (Universität Konstanz)
Gerd Ronning (Universität Tübingen)
Joachim Wagner (Universität Lüneburg)

Data Masking by Noise Addition and the Estimation of Nonparametric Regression Models

By Sandra Lechner and Winfried Pohlmeier, Konstanz*

JEL C21, J24, J31

Data masking, errors-in-variables, SIMEX, local polynomial regression.

Summary

Data collecting institutions use a large range of masking procedures in order to protect data against disclosure. Generally, a masking procedure can be regarded as a kind of data filter that transforms the true data generating process. Such a transformation severely affects the quality of the data and limits its use for empirical research. A popular masking procedure is noise addition, which leads to inconsistent estimates if the additional measurement errors are ignored.

This paper investigates to what extent appropriate econometric techniques can obtain consistent estimates of the true data generating process for parametric and nonparametric models when data is masked by noise addition. We show how the reduction of the data quality can be minimized using the local polynomial Simulation-Extrapolation (SIMEX) estimator. Evidence is provided by a Monte-Carlo study and by an application to firm-level data, where we analyze the impact of innovative activity on employment.

1. Introduction

An increasing demand for microdata can be witnessed among empirical researchers in the last decades. Very often such data contains highly sensitive information, whose confidentiality has to be protected against disclosure by the data collecting institution, not only for legal reasons, but also in order to increase the reliability of the responses by guarantying a maximum amount of confidentiality and privacy to the individual respondents. Hence, for the data collecting institution, masking procedures to protect data against disclosure of the observational units create a trade-off between the goal of providing a maximum amount of information to the empirical researcher and the (often legal) requirement of data protection.

Statistical offices and other data collecting institutions use a wide range of masking procedures in order to avoid or to minimize the probability of disclosing sensitive individual

* This paper is a revised version of our paper presented at the workshop "Econometric Analysis of Anonymised Firm Data", University of Tübingen, March 18th – 19th, 2004. The idea for this paper arose from our work for the scientific advisory board "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" (Actual Anonymization of Individual Economic Data) of the German Federal Statistical Office. We like to thank the Zentrum für Europäische Wirtschaftsforschung (ZEW) for providing us with the data and Gerd Ronning for helpful comments on an earlier version of the paper. Financial support by the DFG is gratefully acknowledged. The usual disclaimer applies.

information.[1] Masking procedures are mainly chosen because they are straightforward and guarantee a high degree of protection against disclosure. The quality of masked data for empirical research is questionable if not appropriate measures are taken.

From a statistical point of view, a masking procedure simply represents a data filter that transforms the true data generating process. Often masking procedures destroy the stochastic structure of the data so that no valid inference can be made. For instance, this holds for the "data-swapping" method, where sensitive information is permuted across units. Such a method implies the destruction of the true correlation structure between covariates, so that any causal analysis becomes infeasible. Thus, for the empirical researcher who is interested in the true data generating process (e.g. the parameters of an econometric model), the obvious question arises to what extent a masking procedure contaminates the true data generating process, and whether specific econometric techniques exist which allow one to infer the true data generating process on the basis of the masked data.

Finally, even if the true data generating process can be estimated consistently by some appropriate econometric technique, the question of efficiency reduction through masking remains open, i.e. masking might reduce the efficiency of the estimates so that inferential statements can only be made conditional on the assumption that the loss of information due to masking did not severely affect the confidence intervals.

This paper analyzes the questions raised above for the case of data protection by noise addition as one of the most popular disclosure techniques. By adding independent errors to the covariates, the method is easy-to-implement. Moreover, from a statistician's point of view, this way of contaminating the original data can be easily understood and it creates nothing but a standard errors-in-variables-problem. For the linear regression model, the effects of measurement errors on the properties of linear estimators are well understood and discussed in the literature on errors-in-variables models (see e.g. Fuller (1987)). In their study on alternative masking procedures, Lechner/Pohlmeier (2003) compare various methods of masking by noise addition with masking by listwise micro-aggregation and their consequences for the linear regression model.

The problem of errors-in-variables in nonlinear models has been tackled in the literature less frequently. Notable exceptions are the papers by Amemiya (1985), Hausmann/Newey/Powell (1995), Lee/Sepanski (1995), Schennach (2004) and Hong/Tamer (2002), which deal with special aspects of measurement errors in nonlinear settings. The monograph by Carroll/Ruppert/Stefanski (1995) surveys various approaches to errors-in-variables in nonlinear models.

Cook/Stefanski (1994) introduce the simulation-extrapolation method (SIMEX), a simulation-based method that can be used for the estimation of nonlinear models with measurement errors, if additional information on the variance of the measurement error is available. Lechner/Pohlmeier (2004) show how this method can be modified for the case of data masking by noise addition. They provide Monte-Carlo evidence for parametric nonlinear models where the SIMEX method is able to substantially mitigate the estimation bias generated by noise addition. From a practical point of view this method is attractive for two reasons: first, SIMEX can be applied for a wide range of popular nonlinear models such as Probit, Logit, Tobit and Poisson regression models, secondly, it rests on a masking procedure that can be implemented by the data collecting institution without any increase of disclosure risk. The data collecting institution only has to provide the empirical researcher with information on the second moments of the error process

[1] See *Domingo-Ferrer/Torra* (2002) or *Brand* (2000) for a review of these procedures.

that is used for the noise addition. Such additional information does not substantially increase the disclosure risk.

Few papers consider the problem of measurement errors in nonparametric regression functions. To our knowledge this problem has been only addressed by three studies. Fan/Truong (1993) propose a globally consistent nonparametric regression estimator based on a deconvolution kernel method. Carroll/Maca/Ruppert (1999) use regression splines to estimate the conditional mean function. This paper builds on the work by Staudenmayer/Ruppert (2004), who propose the SIMEX method for nonparametric regression models with measurement errors. We show that the SIMEX method can be easily modified to the case where data has been masked by noise addition.

The paper is organized as follows. Using the bivariate linear regression model as an example, we introduce in Section 2 the key idea of the SIMEX method in the context of data masking by noise addition. The generalization to the nonparametric setting is presented in Section 3. Section 4 presents evidence on the power of this approach. Based on the results of a Monte-Carlo experiment as well as on an illustrative empirical analysis of the impact innovative activity on employment at the firm-level, we show that the SIMEX method nicely corrects for the estimation bias introduced by data masking through noise addition. Section 5 concludes and gives an outlook on future research.

2. SIMEX and data masking by noise addition

The SIMEX method is a simulation based method of estimating and reducing the bias due to measurement error. Although originally designed for the case of one additive measurement error, it can be extended to measurement errors in several covariates as well as to the case of multiplicative measurement errors.[2] SIMEX is a two-step estimation procedure consisting of a simulation step and an extrapolation step. The key idea is to use the information from an incremental addition of measurement errors with the mismeasured (masked) data W_i using computer simulated random errors. Adding additional measurement error to the data by the simulation exercise allows the statistician to infer in which way the estimation bias is affected by the increase of the variance of the measurement error. This is the so-called simulation step. In the extrapolation step, the estimated parameters are modelled as a function of the magnitude of the variance of the measurement error and extrapolated to the case of no measurement error.

In order to clarify the basic idea of the SIMEX method, we focus on the bivariate linear regression model with additive measurement error. Consider the following linear regression model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1, \ldots, N, \tag{1}$$

where Y_i is the dependent variable, X_i is the explanatory variable, and ε_i is an independent error term, with mean $\mathrm{E}\left[\varepsilon_i | X_i\right] = 0$ and $\mathrm{V}[\varepsilon_i | X_i] = \sigma_\varepsilon^2(X_i)$.

Suppose that the explanatory variable X_i contains sensitive information, which should be protected against disclosure. Rather than observing X_i, we observe a masked explanatory variable W_i defined as:

$$W_i = X_i + u_i, \tag{2}$$

[2] See *Carroll* et al. (1995)