

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.

* 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.