11 Microeconometric Models and Anonymized Micro Data *

Gerd Ronning[1]

[1] Wirtschaftswissenschaftliche Fakultät, Universität Tübingen
gerd.ronning@uni-tuebingen.de

Summary: The paper first provides a short review of the most common microeconometric models including logit, probit, discrete choice, duration models, models for count data and Tobit-type models. In the second part we consider the situation that the micro data have undergone some anonymization procedure which has become an important issue since otherwise confidentiality would not be guaranteed. We shortly describe the most important approaches for data protection which also can be seen as creating errors of measurement by purpose. We also consider the possibility of correcting the estimation procedure while taking into account the anonymization procedure. We illustrate this for the case of binary data which are anonymized by 'post-randomization' and which are used in a probit model. We show the effect of 'naive' estimation, i. e. when disregarding the anonymization procedure. We also show that a 'corrected' estimate is available which is satisfactory in statistical terms. This is also true if parameters of the anonymization procedure have to be estimated, too.

11.1 Introduction

Empirical research in economics has for a long time suffered from the unavailability of individual 'micro' data and has forced econometricians to use (aggregate) time series data in order to estimate, for example, a consumption function. On the contrary other disciplines like psychology, sociology and, last not least, biometry have analyzed micro data already for decades. Therefore it is not surprising that most of the by now well-known microeconometric methods have been invented long time ago by biometricians and psychometricians. However, it is the merit of econometricians

*Research in this paper is related to the project "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" financed by German Ministry of Research and Technology.

that they have provided the underlying behavioral or structural model. For example, the probit model can be seen as an operational version of a linear model which explains the latent dependent variable describing, say, the unobservable reservation wage. Moreover, the discrete choice model results from the hypothesis that choice among alternatives is steered by maximization of utility of these alternatives.

The software for microeconometric models has created growing demand for micro data in economic research, in particular data describing firm behavior. However, such data are not easily available when collected by the Statistical Office because of confidentiality. On the other hand these data would be very useful for testing microeconomic models. This has been pointed out recently by KVI commission.[1] Therefore, the German Statistical Office initiated research on the question whether it is possible to produce scientific use files from these data which have to be anonymized in a way that re-identification is almost impossible and, at the same time, distributional properties of the data do not change too much. Published work on anonymization procedures and its effects on the estimation of microeconometric models has concentrated on *continuous* variables where a variety of procedures is available. See, for example, Ronning and Gnoss (2003) for such procedures and the contribution by Lechner and Pohlmeier (2003) also for the effects on estimation. Discrete variables, however, mostly have been left aside in this discussion. The only stochastic-based procedure to anonymize discrete variables is post-randomization (PRAM) which switches categories with prescribed probability. In this paper we consider anonymization by PRAM and its effect on the estimation of the microeconometric probit model. Thus, we consider an anonymized binary variable which is used as dependent variable in a probit model whereas the explanatory variables remain in their original form.

In Section 11.2 we describe the most important microeconometric models which by now should be well known so that details on estimation and testing are omitted and only principles of modelling are sketched. Section 11.3 presents anonymization procedures and possible strategies to incorporate them into the estimation of microeconometric models. Finally Section 11.4 will illustrate these general remarks for the special case that the binary dependent variable in a probit model has been anonymized by PRAM. We also consider the case that the user is not informed about details of the anonymization procedure.

11.2 Principles of Microeconometric Modelling

Consider the following linear model:

$$Y^* = \alpha + \beta x + \varepsilon \tag{11.1}$$

with $E[\varepsilon] = 0$ and $V[\varepsilon] = \sigma_\varepsilon^2$. Here the $*$ indicates that the continuous variable Y is latent or unobservable. This model asserts that the conditional expectation of Y^* but not the corresponding conditional variance depends on x. If the dependent

[1] See Kommission zur Verbesserung der statistischen Infrastruktur (2001).