WP. 26
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (iv): Access to business microdata for analysis

ESTIMATION OF THE PROBIT MODEL FROM ANONYMIZED MICRODATA

Invited Paper

Submitted by the University of Tuebingen, Germany[1]

[1] Prepared by Gerd Ronning and Martin Rosemann.

Estimation of the Probit Model From Anonymized Micro Data.

Gerd Ronning[*], Martin Rosemann[**]

[*] Department of Economics, University of Tuebingen, Mohlstrasse 36, D-72074 Tuebingen. (gerd.ronning@uni-tuebingen.de)

[**] Institute for Applied Economic Research (IAW), Ob dem Himmelreich 1, D-72074 Tuebingen, (martin.rosemann@iaw.edu)

Abstract. The demand of scientists for confidential micro data from official sources has created discussion of how to anonymize these data in such a way that they can be given to the scientific community. We report results from a German project which exploits various options of anonymization for producing such "scientific-use" files. The main concern in the project however is whether estimation of stochastic models from these perturbed data is possible and - more importantly - leads to reliable results. In this paper we concentrate on estimation of the probit model under the assumption that only anonymized data are available. In particular we assume that the binary dependent variable has undergone post-randomization (PRAM) and that the set of explanatory variables has been perturbed by addition of noise. We employ a maximum likelihood estimator which is consistent if only the dependent variable has been anonymized by PRAM. The errors-in-variables structure of the regressors then is handled by the simulation extrapolation (SIMEX) estimation procedure.

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. The software for microeconometric models has created growing demand for micro data in economic research, in particular data describing firm behaviour. 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. Results from this project have been published quite recently. See Ronning et al. (2005) where most known anonymization procedures have been rated both with regard to data protection and to informational content left after perturbation. In particular we found (rank) swapping procedures not acceptable from user's point of view.

Published work on anonymization of micro data 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 when anonymizing data either by microaggregation or addition of noise. 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 concentrate on estimation of the probit model for which only anonymized data are available. In particular we assume that the binary dependent variable has undergone post-randomization (PRAM) and that the set of explanatory variables has been perturbed by addition of noise. We employ a maximum likelihood estimator which is consistent if only the dependent variable has been anonymized by PRAM. The errors-in-variables structure of the regressors then is handled by the simulation extrapolation (SIMEX) estimation procedure.

In Section 2 we consider the probit model. We assume that the binary dependent variable has been anonymized by PRAM whereas right-hand regressor variables have been left in original form. Consistent estimates are available from an adapted estimation procedure. We then turn to the situation that the continuous regressors have been anonymized by noise addition (section 3). An attractive procedure for handling such situations is the simulation extrapolation (SIMEX) estimator which will be briefly described. Section 4 then presents some estimation results for the probit model when both the dependent and the independent variables have been anonymized. We present results from a simulation study where

[1] See KVI (2001).

the PRAM adapted probit estimator is combined with the SIMEX approach.

2 The probit model under post randomization

2.1 The probit model

Consider the following linear model:[2]

$$Y^* \quad = \quad \alpha \, + \, \beta \, x \, + \, \varepsilon \tag{1}$$

with $E[\varepsilon] \;=\; 0$ and $V[\varepsilon] \;=\; \sigma^2_\varepsilon$. 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. However we observe only a binary variable Y which is related to the latent variable by the "threshold model":

$$Y \quad = \quad \begin{cases} 0 & \text{if } Y^* \leq \tau \\ 1 & \text{else} \quad . \end{cases} \tag{2}$$

It can be shown that two of the four parameters $\alpha, \beta \; \sigma^2_\varepsilon$ and τ have to be fixed in order to attain identification of the two remaining ones. Usually we set $\tau = 0$ and $\sigma^2_\varepsilon = 1$ assuming additionally that the error term ε is normally distributed. This is the famous probit model. Note that only the probability of observing $Y = 1$ for a given x can be determined. If we alternatively assume hat the error term follows a logistic distribution, we obtain the closely related binary logit model.

2.2 Randomized response and post randomization

Randomization of the binary variable Y can be described as follows: Let Y^m denote the 'masked' variable obtained from post randomization. Then the transition probabilities can be defined by $p_{jk} \;\; \equiv \;\; P(Y^m \; = \; j \,|\, Y \; = \; k)$ with $j, k \; \varepsilon \; \{0, 1\}$ and $p_{j0} + p_{j1} = 1$ for $j = 0, 1$. If we define the two probabilities of no change by $p_{00} \equiv \pi_0$ and $p_{11} \equiv \pi_1$, respectively, the probability matrix can be written as follows:

$$\mathbf{P}_y \quad = \quad \begin{pmatrix} \pi_0 & 1 \, - \, \pi_0 \\ & \\ 1 \, - \, \pi_1 & \pi_1 \end{pmatrix}$$

[2] See, for example, Ronning (1991).

Since the two probabilities of the post randomization procedure usually are known and there is no argument not to treat the two states symmetrically, in the following we will consider the special case

$$\pi_0 \quad = \quad \pi_1 \qquad . \tag{3}$$

When the variable Y has undergone randomization, we will have a sample with n observations y_i^m where y_i^m is the dichotomous variable obtained from y_i by the randomization procedure.

In the handbook on anonymization (Ronning et al 2005) we also discuss the extension of PRAM to more than two categories. If the categories are ordered as, for example, in the case of ordinal variables or count data, switching probabilities for adjoining categories should be higher since otherwise the ordering would be totally destroyed. Of course, PRAM could also be extended to joint anonymization of two or more discrete variables.

2.3 Estimation of the model under PRAM

Under randomization of the dependent observed variable we have the following data generating process:

$$Y_i^m \quad = \begin{cases} 1 & \text{with probability} \quad \Phi_i \, \pi \; + \; (1 - \Phi_i) \, (1 - \pi) \\ 0 & \text{with probability} \quad \Phi_i \, (1 - \pi) \; + \; (1 - \Phi_i) \, \pi \end{cases} \tag{4}$$

Here Φ_i denotes the conditional probability under the normal distribution that the unmasked dependent variable Y_i takes on the value 1 for given x_i, i.e. $\Phi_i \equiv \Phi(\alpha + \beta x_i) = P(Y_i^* > 0 \mid x_i)$.

From (4) we obtain the following likelihood function:

$$\begin{aligned} &\mathcal{L}(\alpha, \beta | (y_i^m, x_i), i = 1, \ldots, n) \\ &\quad = \prod_{i=1}^{n} \left(\Phi_i \, \pi \; + \; (1 - \Phi_i) \, (1 - \pi)\right)^{y_i^m} \; \left(\Phi_i \, (1 - \pi) \; + \; ((1 - \Phi_i) \, \pi\right)^{(1 - y_i^m)} \quad . \end{aligned} \tag{5}$$

Global concavity of this function with respect to α and β may be checked by deriving first and second (partial) derivatives of the log-likelihood function. Ronning (2005) derives the Hessian matrix of partial derivatives. A simple formula for the information matrix can be derived from which it is immediately apparent that maximum likelihood estimation under randomization is consistent but implies an efficiency loss which is greatest for values of π near 0.5. See Ronning (2005) for detailed results.

3 Addition of noise and the simulation extrapolation approach

3.1 Data protection by addition of noise

Consider the linear model which we write in usual way as follows: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. Let $\mathbf{e}_y$ be a vector of errors with expectation zero and positive variance corresponding to $\mathbf{y}$ and let $\mathbf{E}_X$ be a matrix of errors corresponding to $\mathbf{X}$. Addition of noise means that we have to estimate the unknown parameter vector from the model

$$\mathbf{y} + \mathbf{e}_y \;=\; (\mathbf{X} + \mathbf{E}_X)\,\boldsymbol{\beta} \;+\; \mathbf{u}\,. \tag{6}$$

This is the well-known errors-in-variables model for which anonymization of right-hand variables creates estimation problems whereas anonymization of the dependent variable only increases the error variance[3] which should be compared with the case of microaggregation where (separate) anonymization of the dependent variable creates problems. Lechner and Pohlmeier (2005) consider nonparametric regression models where the regressors are anonymized by addition of noise. They show that from the simulation-extrapolation method (SIMEX) reliable estimates can be obtained. However for microeconometric models such as logit and probit models general results regarding the effect of noise addition and the suitability of the SIMEX method are not yet established.

Additive errors have the disadvantage that greater values of a variable are less protected. Take as an example sales of firms. If one firm has sales of 1 million and another sales of 100 million then addition of an error of 1 doubles sales of the first but leaves nearly unchanged sales of the second firm. Therefore research has been done also for the case of multiplicative errors which in this case should have expectation one. Formally this leads to

$$\mathbf{y} \odot \mathbf{e}_y \;=\; (\mathbf{X} \odot \mathbf{E}_X)\,\boldsymbol{\beta} \;+\; \mathbf{u}$$

where $\odot$ denotes element-wise multiplication (Hadamard product). For results regarding estimation of this linear model see Ronning et al (2005). In the following we consider only the additive case.

3.2 The SIMEX approach

We will only sketch the idea of this approach[4] for the simple linear regression model which is a special case of the linear model considered above with only one

[3] See Lechner and Pohlmeier (2003) for details.

[4] For details see, for example, Carroll et al (1995).

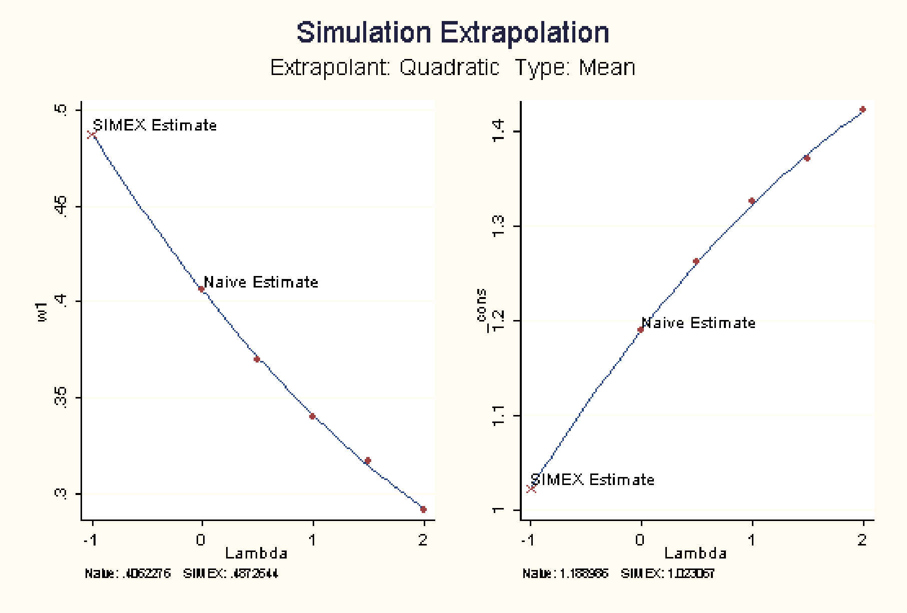

Figure 1: SIMEX estimator – quadratic extrapolation function

regressor and a constant term. It is well known from econometrics that estimation of the regression coefficient β by least squares leads to

$$plim\,\hat{\beta} \;=\; \beta\,\frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2}. \tag{7}$$

if the regressor variable x can only be observed with error e_x where σ_x^2 is the variance of x and σ_e^2 is the variance of this error. This corresponds to equation (6) with $P\left[\mathbf{e}_y = \mathbf{0}\right] = 1$. Now assume that this variance is known and that another error $\lambda\, e_x$ with $\lambda > 0$ is added to the error affected regressor variable by purpose. Then we obtain

$$plim\,\hat{\beta}(\lambda) \;=\; \beta\,\frac{\sigma_x^2}{\sigma_x^2+(1+\lambda)\sigma_e^2} \tag{8}$$

so that a consistent estimator would be obtained for $\lambda = -1$. Of course $\hat{\beta}(\lambda)$ can be evaluated for any positive λ using simulation whereas results for $\lambda < 0$ have to be guessed. Usually M simulation runs are averaged for each λ so that

$$\overline{\hat{\beta}(\lambda)} = \frac{1}{M}\sum_{j=1}^{M}\hat{\beta}_j(\lambda)$$

is the estimate actually used. Cook and Stefanski (1994) suggested an extrapolation procedure which fits a curve to the various points and extrapolates it for $\lambda = -1$. This is illustrated in figure 1 for the case of a quadratic extrapolation function showing results for both coefficients from the simple regression model.

4 Simulation results

In this section we will estimate the two parameters α and β of the probit model defined in (1) and (2) assuming that the dependent variable y has been anonymized by PRAM and that the regressor variable x has been protected by addition of noise. We also assume that the PRAM parameter π and the error variance σ_x^2 are known.[5] Simulated data will be used for estimation.[6]

We assume that $n = 500$ observations are available and that the two unknown parameters are given by $\alpha = -2.5$ and $\beta = 0.6$. The regressor variable is generated from a normal distribution $N(4.35\,; 1.75^2)$ and the error variable satisfies $\varepsilon \sim N(0; 1)$ the latter recognizing the identification constraint of the probit model. The maximum likelihood (ML) estimator of the probit model based on the likelihood function (4) is evaluated by a GAUSS programme written by the first author using the subroutine MAXLIK from the GAUSS library.[7]

We use $R = 50$ iterations in our simulation study which may be considered as too small but was chosen to keep computing time within acceptable limits. In each iteration the ML estimator of the probit model is employed in the SIMEX procedure: First for each $\lambda \, \varepsilon \, \{0\,,\, 0.5\,,\, 1.0\,,\, 1.5\,,\, 2.0\}$ we computed $M = 250$ values of this estimator from which $\overline{\hat{\beta}(\lambda)}$ was determined. Using the five different estimates we then fitted a quadratic function to these five points and obtained the final estimate of both α and β from evaluating this function at $\lambda = -1$. From the $M = 50$ estimates we computed mean, standard deviation, median and both the minimal and the maximal value which are presented in the following table.

[5] It is possible to extend the estimation procedure to the case that π is unknown. See Ronning (2005).

[6] The same design has been used in Ronning et al (2005) where only the dependent variable was anonymized.

[7] Many thanks to Sandra Lechner for providing us with a SIMEX routine!

Table 1: Probit model - PRAM adapted ML and SIMEX procedure

π		estimate	stand.dev.	variance	minimum	median	maximum
1,000	α	-3.581985	0.379676	0.144154	-4.417265	-3.555691	-2.880216
	β	0.860513	0.090857	0.008255	0.720343	0.846306	1.061476
0.975	α	-3.323051	0.368318	0.135658	-4.386024	-3.340966	-2.524121
	β	0.799967	0.088630	0.007855	0.641313	0.786402	1.046858
0,950	α	-3.005847	0.342184	0.117090	-3.870007	-2.980716	-2.274590
	β	0.726057	0.076550	0.005860	0.560696	0.728072	0.889345
0.925	α	-2.750990	0.309059	0.095518	-3.794785	-2.772832	-2.188305
	β	0.660091	0.070754	0.005006	0.532053	0.660100	0.919395
0,900	α	-2.498118	0.257171	0.066137	-3.073287	-2.505895	-2.013187
	β	0.597422	0.051819	0.002685	0.485514	0.600677	0.704354
0.875	α	-2.304071	0.200737	0.040295	-2.994306	-2.280229	-1.982457
	β	0.553141	0.049268	0.002427	0.473322	0.545844	0.700764
0,850	α	-2.051473	0.207291	0.042970	-2.791021	-2.043950	-1.729020
	β	0.488270	0.041388	0.001713	0.422237	0.490443	0.632154
0.825	α	-1.789513	0.190780	0.036397	-2.349686	-1.819519	-1.398894
	β	0.431177	0.040328	0.001626	0.349578	0.434013	0.549036
0,800	α	-1.543171	0.136212	0.018554	-1.882205	-1.530583	-1.272405
	β	0.372731	0.029233	0.000855	0.310063	0.369021	0.451543
Remarks: Simulation design: $\alpha = -2.5, \beta = 0.6, \sigma_e^2 = 0,01, n = 500, R = 50, M = 250$							

Since we know from earlier simulation experiments that values of the PRAM parameter π create computational problems if π is far away from 1.0 we confined simulation to the interval $\pi \; \varepsilon \; [0.8\,; 1.0]$. Noise addition is done by a normally distributed variable with $\sigma_e^2 = 0.01$. We plan to vary both π and σ_e^2 in a systematic manner since we found that estimation results seem to be very sensitive to combinations of these two parameters chosen. See the simulation results.

The table shows that for $\pi = 1.00$ (no post randomization) both parameters show a bias "away from zero" which becomes smaller and switches its sign for decreasing values of π. Whether this is only a effect of the special type of the extrapolation function has to be analyzed in more detail.

References

Carroll, R.J. Ruppert, D., and Stefanski, L.A., (1995) *Measurement Error in Nonlinear Models*, Chapman and Hall, London

Cook, J.R. and Stefanski, L.A.(1994) "Simulation-Extrapolation Estimation in Parametric Measurement Error Models".*Journal of the American Statistical Association* **89**, 1314–1328.

Kommission zur Verbesserung der informationellen Infrastruktur (editor) (2001). *Wege zu einer besseren informationellen Infrastruktur.* Nomos, Wiesbaden , cited as KVI(2001).

Lechner, and S., Pohlmeier, W. (2003) "Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten" In: Gnoss, R. und G. Ronnning (editors) *Anonymisierung wirtschaftsstatistischer Einzeldaten.* Forum der Bundesstatistik, volume 42, 115-137.

Lechner, and S., Pohlmeier, W. (2005) "Data Masking by Noise Addition and the Estimation of Nonparametric Regression Models"'.*Jahrbücher für Nationalökonomie und Statistik* **225**, 517-528.

Ronning, G. (1991). *Mikroökonometrie.* Springer, Berlin.

Ronning, G. (2005) "Randomized response and the binary probit model". *Economics Letters* **86**, 221-228.

Ronning, G., Gnoss, R. (Editors) (2003) *Anonymisierung wirtschaftsstatistischer Einzeldaten.* Statistisches Bundesamt. Forum der Bundesstatistik, volume 42, Wiesbaden.

Ronning, G., Sturm, R., Höhne, J., Lenz, J., Rosemann, M., Scheffler, M., Vorgrimler, D. (2005) *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten.* Statistisches Bundesamt, Wiesbaden , Reihe "Statistik und Wissenschaft", volume 4.