Linear Models With Measurement Errors Arising From Mixture Distributions

Gerd Ronning, Universität Tübingen, Department of Economics

The 8th Tartu Conference on MULTIVARIATE STATISTICS, 26-29 June 2007 Tartu, ESTONIA

Outline of my talk:

- Scientific-use files from entrepreneurial micro data: A German project
- Multiplicative noise and mixture distributions - a new approach
- Effect on the estimation of linear panel models
- Derivation of consistent estimators using knowledge about anonymization procedure

Scientific-use files from entrepreneurial micro data:

- German law asks for "factual anonymization" of micro data from The German Statistical Office before they can be released
- Protection of data from firms should be higher than for persons
- Anonymization of micro data by
  - microaggregation or
  - addition of stochastic noise (continuous variables)
  - "Post Randomization" (PRAM) (discrete variables)
  - Multiple imputation (suggested by D. Rubin)

.

Scientific-use files from entrepreneurial micro data:

- One aim of the project is the study of the behaviour of estimators when applied to anonymized micro data
- and the correction of the estimation procedure if necessary and if possible.
- Example:
  - A binary dependent variable is anonymized by PRAM and some regressors by addition of stochastic noise.
  - From the PRAM-corrected likelihood of the probit model a ML estimator is derived which is used in a SIMEX procedure to take account of "measurement" errors in the regressors.
- See "Estimation of the Probit Model From Anonymized Micro Data" by Ronning and Rosemann (2006)

Estimation of linear model from anonymized panel data:

- Anonymization of panel data as a new challenge after having first considered only cross-section data
- In this talk: Estimation of a linear panel model from micro data
- which have been anonymized by multiplicative noise
- where noise is generated from a (bimodal) mixture distribution.
- Since all variables are treated jointly, a *multivariate* mixture distribution is involved.

True model (only one regressor to keep it simple):

$$y_{it} \quad = \quad \alpha + \beta\, x_{it} + \tau_i + \eta_{it} \quad , \; i = 1, \ldots, n \; , \; t = 1, \ldots T$$

Observed variables:

$$y^a_{it} = y_{it} \cdot u_{ity} \quad \textbf{and} \quad y^a_{it} = y_{it} \cdot u_{ity}$$

where noise u is generated from

$$u_{ity} = 1 \; + \; \delta\, D_i \; + \; \varepsilon_{ity}$$

and

$$u_{ity} = 1 \; + \; \delta\, D_i \; + \; \varepsilon_{ity}$$

and the binary random variable D_i is given by

$$D \; = \; \begin{cases} +1 & \textbf{with probability 0.5} \\ -1 & \textbf{with probability 0.5} \end{cases}$$

Remarks:

- Both u_{ity} and u_{itx} have expected value of 1 so that original and anonymized variables have the same mean.
- Both u_{ity} and u_{itx} depend on the same 'factor' D_i for a certain point of time which of course induces correlation between these two errors!
- It can be shown that these error specifications are equivalent to the assumption that both errors are generated jointly from a multivariate (in this case: bivariate) mixture distribution.
- The two components of the mixture distribution may follow, for example, a normal distribution or a lognormal distribution. The latter is preferred in case of multiplicative noise.

Why mixture distributions and why errors with factor structure ?

- Bimodal error distributions intensify the protection of data since only little probability mass of the error distribution is located around the 'original' value.
- The factor structure is used in order to preserve proportional relations between y and x:
  y^a/x^a should have a distribution which approximates the distribution of y/x!
- To see this compare

$$\frac{y_{it}}{xit} \qquad \text{and} \qquad \frac{y_{it}^a}{x_{it}^a} = \frac{y_{it} \cdot (1 \; + \; \delta \, D_i \; + \; \varepsilon_{ity})}{x_{it} \cdot (1 \; + \; \delta \, D_i \; + \; \varepsilon_{itx})}$$

  for "small" ε's.

Remarks on the use of the factor structure

- Preservation of proportionality is considered as important in descriptive statistics, at least by most economists.
- Some computational examples show that the factor structure will not change the ratio y/x considerably when y^a/x^a is used instead.
- However: The ratio will be biased already for the original variables since for the expected value of the quotient we obtain

$$E[\frac{Y}{X}] \approx \underbrace{\frac{E[Y]}{E[X]}}_{'true'quotient} + \underbrace{\frac{E[Y]}{(E[X])^3}V[X] - \frac{1}{(E[X])^2}cov[Y,X]}_{bias}$$

Estimation from anonymized panel data:

- **The 'naive' estimator ("within estimator")**

$$\hat{\beta}^a_W \quad = \quad \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it} - \overline{x}^a{}_{i\bullet})(y^a_{it} - \overline{y^a}_{i\bullet})}{\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it} - \overline{x}^a{}_{i\bullet})^2}$$

with

$$\overline{x}^a{}_{i\bullet} \;=\; \frac{1}{T}\sum_{t=1}^{T} x^a_{it} \quad , \quad \overline{y^a}_{i\bullet} \;=\; \frac{1}{T}\sum_{t=1}^{T} y^a_{it} \qquad .$$

- **is not consistent:**

$$\mathbf{plim}_{N\to\infty}(\hat{\beta}^a_W) \;=\; \frac{(1\,+\,\delta^2)\,\sigma_x^2\,\beta}{(1+\delta^2)\,\sigma_x^2 \;+\; \sigma_\varepsilon^2\,(\sigma_x^2 \,+\, \mu_x^2)} \quad \neq \quad \beta$$

except for the case $\delta = 0\,,\ \sigma_\varepsilon^2 = 0.$

- **Since δ and σ_ε^2 are known (if released by the Statistical Office!) and both σ_x^2 and μ_x^2 can be estimated, a corrected estimator can be obtained which will be consistent.**
- **In the case considered here we get**

$$\hat{\beta}_W^{a,corrected} = \frac{(1+\delta^2)\,\widehat{\sigma_x^2} + \sigma_\varepsilon^2\,(\widehat{\sigma_x^2} + \widehat{\mu_x}^2)}{(1+\delta^2)\,\widehat{\sigma_x^2}}\ \hat{\beta}_W^a \quad .$$

with

$$\widehat{\sigma_x^2} = \frac{s_{x^a}^2 - (\delta^2 + \sigma_\varepsilon^2)\,\overline{x}^{a2}}{1 + \delta^2 + \sigma_\varepsilon^2} \quad ,$$

and

$$\widehat{\mu_x} \quad = \quad \overline{x}^a \quad = \quad \frac{1}{nT} \sum_t \sum_i x_{it}^a$$

Remarks:

- Instead of the multiplicative one could use an additive version:

$$y_{it}^{a} = y_{it} + u_{it}$$

where

$$u_{it} = \mu\, D_i \; + \; \varepsilon_{it}$$

- Although this case is easier to treat analytically we prefer the multiplicative version since it protects more efficiently large firms.
- Results regarding bias change if only regressors are anonymized by stochastic noise. However, this case is unlikely since all variables will be anonymized jointly.
- The estimator is consistent if only the dependent variable is anonymized (either additively or multiplicatively).

Open questions and further research:

- Results for non-linear panel models (probit and logit models, Tobit model, duration models)
- The SIMEX procedure first suggested by Carroll, Ruppert and Stefanski (1995) could be applied.
- However, in case of correlated errors some modifications are necessary.
- In particular this is important if the errors regarding the dependent variable are correlated with errors regarding the regressors.

Some references:

Ronning, G. et al. (2005). Handbook of Anonymization of Entrepreneurial Micro Data. (*German*). Statistisches Bundesamt, Wiesbaden , Reihe "Statistik und Wissenschaft", Band 4, 2005.

Ronning, G. and M. Rosemann(2006). Estimation of the Probit Model From Anonymized Micro Data. Manuscript, April 2006, submitted to Journal of Official Statistics.

Ronning, G (2007). Stochastic Noise Using Mixture Distributions .(*German*). IAW Discussion Paper No. 30 (April 2007). IAW, Tübingen.IM

Regarding SIMEX:

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

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.