On Sample Selection Models and Skew Distributions

Emmanuel O. Ogundimu[1] **and Jane L. Hutton**[1]

[1]*Department of Statistics, University of Warwick, UK*

Abstract

Scores arising from questionnaires often follow asymmetric distributions, on a fixed range. This can be due to scores clustering at one end of the scale or selective reporting. Sometimes, the scores are further subjected to sample selection resulting in partial observability. Thus, methods based on complete cases for skew data are inadequate for the analysis of such data and a general sample selection model is required. Heckman proposed a full maximum likelihood estimation method under the normality assumption for sample selection problems, and parametric and non-parametric extensions have been proposed. A general selection distribution for a vector $\mathbf{Y} \in \mathbb{R}^p$ has a PDF f_Y given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}^\star}(\mathbf{y}) \frac{P(\mathbf{S}^\star \in \mathbf{C} \,|\, \mathbf{Y}^\star = y)}{P(\mathbf{S}^\star \in \mathbf{C})},$$

where $\mathbf{S}^\star \in \mathbb{R}^q$ and $\mathbf{Y}^\star \in \mathbb{R}^p$ are two random vectors, and $\mathbf{C}$ is a measurable subset of $\mathbb{R}^q$. We use this generalization to develop a sample selection model with underlying skew-normal distribution. A link is established between the continuous component of our model log-likelihood function and an extended version of a generalized skew-normal distribution. This link is used to derive the expected value of the model, which extends Heckman's two-stage method. Finite sample performance of the maximum likelihood estimator of the model is studied via Monte Carlo simulation. The model parameters are more precisely estimated under the new model, even in the presence of moderate to extreme skewness, than the Heckman selection models. Application to data from a study of neck injuries, where the responses are substantially skew, successfully discriminates between selection and inherent skewness. We also discuss computational and identification issues, and provide an extension of the model using underlying skew-t distribution.

Keywords: Generalized Sample selection, Missing data, Closed Skew-normal distribution, Closed Skew-t distribution.
AMS subject classifications: 62D99

Bibliography

[1] Heckman, J. (1976) . The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, 475–492.

[2] Copas, J. B. and Li, H. (1997). Inference for non-random samples J. R. Statist. Soc. B 59, 55–95.