Efficient Estimation And Model Selection for Grouped Data with Local Moments *

K. Hitomi, Q-F. Liu, Y. Nishiyama, N. Sueishi

December 15, 2007

Abstract

This paper proposes efficient estimation methods for analyzing grouped data when frequencies as well as local moments are available for each group. Assuming the original data is an i.i.d. sample from a parametric density with unknown parameters, we obtain the joint density of frequencies and local moments, and propose a maximum likelihood (ML) estimator. We further compare it with the generalized method of moments (GMM) estimator and prove these two estimators are asymptotically equivalent in the first order. Based on the ML method, we propose to use Akaike information criterion (AIC) for model selection. We also provide a specification test based on the GMM estimation. Monte Carlo experiments show that the estimators perform remarkably well, AIC selects the right model with high frequency, and the specification test has good size and power properties.

Keywords: Grouped data; Local moments; MLE; GMM; Specification test; AIC; Model Selection.

1 Introduction

In practice, some data are provided only in a grouped form. An example is personal income data reported by government organizations. They provide only masked data for confidential reasons. Typically, income distribution is divided into some classes (by age, for instance) and only summary statistics such as frequencies and class-wise means are observable to researchers. Also, we often see insurance claim data in the same form. Researchers cannot directly observe the claim sizes of each accident, but avail only of some summaries for each stratum. Throughout this paper, we assume that the "original" sample

*We would like to thank Atsushi Yoshida, Yoshinori Kawasaki, seminar participants at Kyoto University and participants for JEA annual meeting at Kyoto Sangyo University for helpful comments. This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grand-in-Aid for 21st Century COE Program "Interfaces for Advanced Economic Analysis".

$\{x_i\},\ i = 1, \cdots, n$, which is unavailable for statisticians, is a realization of a random sample $\{X_i\}$, $i = 1, \cdots, n$ from a distribution with parametric density $f(x; \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ is a vector of unknown parameters. We also suppose that the bounds of each stratum are non-random.

A common and classical situation is the case when only the frequencies are available. Suppose that the support of X_1 is divided into a set of fixed disjoint classes $B_1, B_2, \ldots, B_L$, and we observe only the frequency n_j in each of the classes B_j. Since the individual data are not available, the following standard maximum likelihood estimator is infeasible:

$$\hat{\boldsymbol{\theta}}_{iMLE} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}) \tag{1.1}$$

even though we know the explicit form of the density. The subscript iMLE indicates infeasible maximum likelihood estimator. In this case, however, we easily obtain the log-likelihood function with respect to $n_j,\ j = 1, \cdots, L$, which equals to $\sum_{j=1}^{L} n_j \log P_j(\boldsymbol{\theta})$, where $P_j(\boldsymbol{\theta}) = \int_{B_j} f(x; \boldsymbol{\theta}) dx$ is the probability that an observation falls in B_j. This gives an MLE of $\boldsymbol{\theta}$ as a solution to the normal equation:

$$\sum_{j=1}^{L} n_j \frac{\partial \log P_j(\hat{\boldsymbol{\theta}}_{nMLE})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{1.2}$$

We call it the naive MLE (nMLE). Asymptotic properties of nML have been examined in several papers (see, for example, Tallis (1967)). It is consistent for θ_0, the true value of θ, and asymptotically normally distributed with covariance matrix $-\sum_{j=1}^{L} P_j(\theta_0) \frac{\partial^2 \log P_j(\theta_0)}{\partial \theta \partial \theta'}$. Victoria-Feser and Ronchetti (1997) discuss the properties of the nMLE and some related estimators in terms of robustness. An estimator easy to compute is proposed by Brix and Pfeifer (2000) which does not require explicitly evaluating $P_j(\theta)$. See also Yanagimoto (1990) and Wooldridge (2001) which treat related topics.

In this paper, we consider the situation in which local moments in each class, namely,

$$\bar{x}_j^{(k)} = \frac{1}{n_j} \sum_{i=1}^{n} I(x_i \in B_j) x_i^k, \quad i = 1, \cdots, n, \quad k = 1, \cdots, K$$

are also available in addition to the frequencies n_j, where $I(\cdot)$ dnotes the indicator function.

The purpose of this paper is to propose an efficient estimation method by maximum likelihood in this setup and a model selection method, as well as to provide a specication test for assumed form of parametric density. We also show that the GMM estimator also attains the same efficiency as the MLE. Though we believe we can proceed for any integer K satisfying $E(X_1^K) < \infty$; we only discuss the case when $K = 1$: It is obvious that GMM can handle the higher order moments case more straightforwardly.