Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 3368–3380
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Estimation of the income distribution and detection of subpopulations: An explanatory model

Emmanuel Flachaire[a,*], Olivier Nuñez[b]

[a] *CES, Université Paris 1 Panthéon-Sorbonne, 106-112 bd de l'Hopital 75013 Paris, France*
[b] *Universidad Carlos III de Madrid, Calle Madrid 126, 28903 Getafe, Madrid, Spain*

Available online 28 July 2006

Abstract

Empirical evidence, obtained from non-parametric estimation of the income distribution, exhibits strong heterogeneity in most populations of interest. It is common, therefore, to suspect that the population is composed of several homogeneous subpopulations. Such an assumption leads us to consider mixed income distributions whose components feature the distributions of the incomes of a particular homogeneous subpopulation. A model with mixing probabilities that are allowed to vary with exogenous individual variables that characterize each subpopulation is developed. This model simultaneously provides a flexible estimation of the income distribution, a breakdown into several subpopulations and an explanation of income heterogeneity.

Keywords: Income distribution; Mixture models

1. Introduction

In inequality analysis, parametric and non-parametric estimation often suggests heavy-tails or bi-modality in the income distribution (Marron and Schmitz, 1992; Schluter and Trede, 2002; Davidson and Flachaire, 2004). This suggests heterogeneity in the underlying population. To model this heterogeneity it is natural to assume that the population can be broken down into several homogeneous subpopulations. This is the starting point of our paper. Empirical studies on income distribution indicate that the Lognormal distribution fits homogeneous subpopulations quite well (Aitchison and Brown, 1957; Weiss, 1972). And the theory of mixture models indicates that, under regularity conditions, any probability density can be consistently estimated by a mixture of normal densities (see Ghosal and van der Vaart, 2001 for recent results about rates of convergence). Thus, from the relationship between the Normal and Lognormal distributions, we see that any probability density with a positive support (as for instance income distribution) can be consistently estimated by a mixture of Lognormal densities. We expect, then, to be able to estimate closely the true income distribution with a finite mixture of Lognormal distributions and so to identify the subpopulations.

In this paper, we analyse conditional income distributions using Lognormal mixtures. Our contribution is to propose a conditional model by specifying the mixing probabilities as a particular set of functions of individual characteristics. This allows us to characterize the distinct homogeneous subpopulations: we assume that an individual's belonging to a

* Corresponding author. Tel.: +33 14407 8214; fax: +33 14407 8231.
E-mail addresses: emmanuel.flachaire@univ-paris1.fr (E. Flachaire), nunez@est-econ.uc3m.es (O. Nuñez).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.07.004

specific subpopulation can be explained by his individual characteristics. For instance, households with no working adult are more likely to be nearer the bottom of the income distribution than those with all-working adults. The probability of belonging to a given subpopulation, then, may vary among individuals as explained by individual characteristics.

The method is applied to disposable household income, as obtained from a survey studying changes in inequality and polarization in the UK in the 1980s and 1990s. This empirical study demonstrates the usefulness of our method and, although the results are all confirmed by previous studies, they do not lead to conclusions as rich as those achieved here. We find that our method produces results that are readily given to economic interpretation.

The paper is organized as follows: we present our explanatory mixture model in Section 2 and illustrate its use in Section 3.

2. The explanatory mixture model

We assume that the population can be broken down into K homogeneous subpopulations with a proportion p_k of the population, each being a logarithmic-transformation of the Normal distribution with mean μ_k and standard deviation σ_k. Thus, the density function of the income distribution in the population is defined as

$$f(y) = \sum_{k=1}^{K} p_k \Lambda\left(y; \mu_k, \sigma_k\right), \tag{1}$$

where $\Lambda(.; \mu, \sigma)$ is the Lognormal distribution with parameters μ and σ. Note that, as with the number of modes used to detect heterogeneity, the number of components in the mixture is invariant under a continuous and monotonic transformation of income Y. So, if Y is a mixture of K Lognormal densities, then $\log(Y)$ is a mixture of K Normal densities.

A conditional model can be constructed by letting the mixing probabilities vary with exogenous individual characteristics. Given a vector of individual characteristics X, we consider that the income of an individual with these characteristics is distributed according to the mixture

$$f(y\,|X) = \sum_{k=1}^{K} p_k(X) \Lambda\left(y\,; \mu_k, \sigma_k\right), \tag{2}$$

where $p_k(X)$ is the probability of belonging to the homogeneous subpopulation k. We can typically assume that these mixing probabilities depend on a linear index of X. Note that this model is more flexible than the classical analysis of variance, since the probability of belonging to one subpopulation is not necessarily 1 or 0. Moreover, the range of values of the household characteristics which characterize the subpopulation are not pre-fixed but determined by the sample.

For a fixed number of components K, we can estimate $f(y)$ by maximum likelihood (ML) (Titterington et al., 1985; Lindsay, 1995), and $f(y\,|X)$ with a specific algorithm, the details of which are given below. In practice, the number of components K is unknown and can be chosen as that which minimizes some criterion. There is a large number of criteria and the literature on this subject is still in progress (McLachlan and Peel, 2000). The optimal criterion for our model requires more study, which we leave to future work. For the moment, we select the K that minimizes the BIC criterion (Schwarz, 1978), which is known to give consistent estimation of K in mixture models (Keribin, 2000).

An alternative conditional model could be constructed by assuming the individual characteristics influence the magnitude of the group-specific earnings μ_k. Then, the individual characteristics could be used to model the mean of the subpopulations rather than the probabilities of belonging to a subpopulation. This conditional model could be written as

$$f(y\,|X) = \sum_{k=1}^{K} p_k \Lambda\left(y\,; \mu_k(X), \sigma_k\right), \tag{3}$$

where the conditional mean is typically assumed to depend linearly on X, i.e., $\mu_k(X) = X\beta_k$. Conditioning means is relevant when one wishes to analyse the intra-group variability, whereas conditioning probabilities applies when focusing on inter-group variability. In inequality measurement, the major concern is more often to detect the individual