Abstract

Modelling Lorenz curves (LC) for stochastic dominance comparisons is central to the analysis of income distributions. It is conventional to use non-parametric statistics based on empirical income cumulants which are used in the construction of LC and other related second-order dominance criteria. However, although attractive because of its simplicity and its apparent flexibility, this approach suffers from important drawbacks. While no assumptions need to be made regarding the data-generating process (income distribution model), the empirical LC can be very sensitive to data particularities, especially in the upper tail of the distribution. This robustness problem can lead in practice to “wrong” interpretation of dominance orders. A possible remedy for this problem is the use of parametric or semi-parametric models for the data-generating process and robust estimators to obtain parameter estimates. In this paper, we focus on the robust estimation of semi-parametric LC and investigate issues such as sensitivity of LC estimators to data contamination (Cowell and Victoria-Feser, 2002), trimmed LC (Cowell and Victoria-Feser, 2006), and inference for trimmed LC (Cowell and Victoria-Feser, 2003), robust semi-parametric estimation for LC (Cowell and Victoria-Feser, 2007), selection of optimal thresholds for (robust) semi-parametric modelling (Dupuis and Victoria-Feser, 2006), and use both simulations and real data to illustrate these points.
1 Introduction

The Lorenz curve is central to the analysis of income distributions, embodying fundamental intuition about inequality comparisons (Dagum, 1985; Cowell and Victoria-Feser, 2007). Ranking theorems based on Lorenz dominance and the associated concept of stochastic dominance are fundamental to the theoretical welfare economics of distributions. But formal welfare propositions can only be satisfactorily invoked for empirical constructs if sample data can be taken as a reasonable representation of the underlying income distributions under consideration. In practice income-distribution data may be contaminated by recording errors, measurement errors and the like and, if the data cannot be purged of these, welfare conclusions drawn from the data can be seriously misleading. Indeed, it has been formally shown that Lorenz and stochastic dominance results are non-robust (Cowell and Victoria-Feser, 2002). This means that small amounts of data contamination in the wrong place can reverse unambiguous ranking orders: the “wrong place” usually means in the upper tail of the distribution. This is of particular interest in view of a burgeoning recent literature that has focused on empirical issues concerning the upper tail of both income distributions and wealth distributions (Atkinson, 2004; Kopczuk and Saez, 2004; Moriguchi and Saez, 1991; Piketty, 2001; Piketty and Saez, 2003; Saez and Veall, 2005). So it is important to have an approach that enables one to control for the distortionary effect of upper-tail contamination in a systematic fashion. This paper addresses the problem by introducing a robust method of estimating Lorenz curves and implementing stochastic-dominance criteria. To this end we have assembled some recent research on this issue, mainly drawing on the results of Cowell and Victoria-Feser (2006) and Cowell and Victoria-Feser (2007).

Our approach is organized as follows. We begin, in section 2, by setting out the formal background to the Lorenz curve and the estimation problems associated with extreme values. Section 3 develops the semi-parametric approach to modelling Lorenz curves and section 4 discusses the practical problem of parameter choice in implementing the method. Section 5 applies the method to UK data and section 6 concludes.

2 Background

We may set out the formal representation of the Lorenz curve using the following simple framework. Let $\mathcal{F}$ be the set of all univariate probability distributions and $X$ be a random variable with probability distribution $F \in \mathcal{F}$ and support $\mathcal{X} \subseteq \mathbb{R}$. $F$ can be thought of as a parametric model $F_\theta$. We shall write statistics of any distribution $F \in \mathcal{F}$ as a functional $T(F)$; in particular we write the mean as $\mu(F) := \int x dF(x)$. A key distributional concept derived from $F$ is given by the $q^{th}$ cumulative functional $C : \mathcal{F} \times [0, 1] \mapsto \mathcal{X}$:

$$C(F; q) := \int_\Delta Q(F; q) x dF(x) = c_q.$$  \hspace{1cm} (13.1)