Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 141 (2007) 1044–1072
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Income distribution and inequality measurement: The problem of extreme values

Frank A. Cowell[a,*], Emmanuel Flachaire[b]

[a]*STICERD, London School of Economics, Houghton Street, London WC2A 2AE, UK*
[b]*University of Paris 1, Paris School of Economics, 106-112 bd de l'Hopital, 75013 Paris, France*

Available online 12 February 2007

Abstract

We examine the statistical performance of inequality indices in the presence of extreme values in the data and show that these indices are very sensitive to the properties of the income distribution. Estimation and inference can be dramatically affected, especially when the tail of the income distribution is heavy, even when standard bootstrap methods are employed. However, use of appropriate semiparametric methods for modelling the upper tail can greatly improve the performance of even those inequality indices that are normally considered particularly sensitive to extreme values.

JEL classification: C1; D63

Keywords: Inequality measures; Statistical performance; Robustness

1. Introduction

There is a folk wisdom about inequality measures concerning their empirical performance. Some indices are commonly supposed to be particularly sensitive to specific types of change in the income distribution and may be rejected *a priori* in favour of others that are presumed to be "safer". This folk wisdom is only partially supported by formal analysis and it is appropriate to examine the issue by considering the behaviour of inequality measures with respect to extreme values. An extreme value is an observation

*Corresponding author.
E-mail addresses: f.cowell@lse.ac.uk (F.A. Cowell), emmanuel.flachaire@univ-paris1.fr (E. Flachaire).

0304-4076/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.jeconom.2007.01.001

that is highly influential on the estimate of an inequality measure. It is clear that an extreme value is not necessarily an error or some form of contamination. It could in fact be an informative observation belonging to the true distribution—*a high-leverage observation*. In this paper, we study sensitivity of different inequality measures to extreme values, in both cases of contamination and of high-leverage observations.

What is a "sensitive" inequality measure? This issue has been addressed in ad hoc discussion of individual measures in terms of their empirical performance on actual data (Braulke, 1983). Some of the welfare-theoretical literature focuses on transfer sensitivity (Shorrocks and Foster, 1987) and related concepts. But it is clear that informal discussion is not a satisfactory approach for characterising alternative indices; furthermore, the welfare properties of inequality measures in terms of the relative impact of transfers at different income levels will not provide a reliable guide to the way in which the measures may respond to extreme values. We need a general and empirically applicable tool.

Specifically we need to address four key issues that are relevant to assessing the sensitivity of inequality measures: (1) influence functions and their performance in presence of contamination; (2) sensitivity to high-leverage observations; (3) error in probability of rejection in tests with finite samples; (4) sensitivity under different underlying distributions/shapes of tails. The paper will provide general results and simulation studies for each of these four topics using a variety of common inequality indices, in order to yield methods that are implementable in practice.

In Section 2, we examine the sensitivity of inequality measures to contamination in the data, both in high and low incomes. In Section 3, we study the sensitivity of inequality measures to "high-leverage" observations. We investigate Monte Carlo simulations to study the error in the rejection probability (ERP) of a test in finite samples. Section 4 examines the relationship between the apparent sensitivity of the inequality index and the shape of the income distribution. Section 5 proposes a method for detecting extreme values in practice and Section 6 concludes.

2. Data contamination

The principal tool that we use for evaluating the influence of data contamination on estimates is the influence function (*IF*), taken from the theory of robust estimation. If the *IF* is unbounded for some income value z it means that the estimate of the inequality index may be catastrophically affected by data contamination at z or a value close to z. Cowell and Victoria-Feser (1996) have shown that "if the mean has to be estimated from the sample then all scale independent or translation independent and decomposable measures have an unbounded *IF*" and that inequality measures are typically not robust to data contamination.[1] In this section, we re-examine this negative conclusion by using a more detailed comparison of the sensitivity of inequality measures. First we demonstrate how sensitive various inequality measures are to contamination in high and small incomes (Section 2.1). Then, we propose a semiparametric method of obtaining inequality measures that are much less sensitive to data contamination (Section 2.2).

[1]The issue of data contamination is complementary to, but distinct from, the issue of measurement error that has been addressed by Chesher and Schluter (2002). These two issues typically require different estimation techniques.