Available online at www.sciencedirect.com

ScienceDirect

Physica A 370 (2006) 49–53

PHYSICA A

www.elsevier.com/locate/physa

The power-law tail exponent of income distributions

F. Clementi[a,b,*], T. Di Matteo[b], M. Gallegati[c]

[a] *Department of Public Economics, University of Rome "La Sapienza", Via del Castro Laurenziano 9, 00161 Rome, Italy*

[b] *Applied Mathematics, Research School of Physical Sciences and Engineering, The Australian National University, 0200 Canberra, Australia*

[c] *Department of Economics, Università Politecnica delle Marche, Piazzale Martelli 8, 60121 Ancona, Italy*

Available online 12 May 2006

Abstract

In this paper we tackle the problem of estimating the power-law tail exponent of income distributions by using the Hill's estimator. A subsample semi-parametric bootstrap procedure minimizing the mean squared error is used to choose the power-law cutoff value optimally. This technique is applied to personal income data for Australia and Italy.

Keywords: Personal income; Pareto's index; Hill's estimator; Bootstrap

1. Introduction

Since Pareto it has been recognized that a *power-law* provides a good fit for the distribution of high incomes [1]. The Pareto's law asserts that the complementary cumulative distribution $P_>(y) = 1 - \int_{-\infty}^{y} p(\xi)\,\mathrm{d}\xi \to P_>(u)(u/y)^{\alpha}$, with $y \geqslant u$, where $u > 0$ is the threshold value of the distribution and $\alpha > 0$ turns out to be some kind of index of inequality of distribution. The fit of such distribution is usually performed by judging the degree of linearity in a double logarithmic plot involving the empirical and theoretical distribution functions, in such a way that the estimation of u of the distribution does not seem to follow a neutral procedure. Moreover, recent studies have criticized the reliability of this geometrical method by showing that linear-fit based methods for estimating the power-law exponent tend to provide biased estimates, while the maximum likelihood estimation method produces more accurate and robust estimates [2,3]. Hill proposed a conditional maximum likelihood estimator for α based on the k largest order statistics for non-negative data with a Pareto's tail [4]. That is, if $y_{[n]} \geqslant y_{[n-1]} \geqslant \cdots \geqslant y_{[n-k]} \geqslant \cdots \geqslant y_{[1]}$, with $y_{[i]}$ denoting the ith order statistic, are the sample elements put in descending order, then the Hill's estimator is

$$\hat{\alpha}_n(k) = \left[\frac{1}{k}\sum_{i=1}^{k}(\log y_{n-i+1} - \log y_{n-k})\right]^{-1}, \tag{1}$$

*Corresponding author. Department of Public Economics, University of Rome "La Sapienza", Via del Castro Laurenziano 9, 00161 Rome, Italy. Tel.: +39 06 49766843; fax: +39 06 4461964.

E-mail address: fabio.clementi@uniroma1.it (F. Clementi).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2006.04.027

where n is the sample size and k an integer value in $[1, n]$. Unfortunately, the finite-sample properties of the estimator (Eq. (1)) depend crucially on the choice of k: increasing k reduces the variance because more data are used, but it increases the bias because the power-law is assumed to hold only in the extreme tail.

Over the last 20 years, estimation of the Pareto's index has received considerable attention in extreme value statistics [5]. All of the proposed estimators, including the Hill's estimator, are based on the assumption that the number of observations in the upper tail to be included, k, is known. In practice, k is unknown; therefore, the first task is to identify which values are really extreme values. Tools from exploratory data analysis, as the quantile-quantile plot and/or the mean excess plot, might prove helpful in detecting graphically the quantile $y_{[n-k]}$ above which the Pareto's relationship is valid; however, they do not propose any formal computable method and, imposing an arbitrary threshold, they only give very rough estimates of the range of extreme values.

Given the bias-variance *trade-off* for the Hill's estimator, a general and formal approach in determining the best k value is the minimization of the *Mean Squared Error* (*MSE*) between $\hat{\alpha}_n(k)$ and the theoretical value α. Unfortunately, in empirical studies of data the theoretical value of α is not known. Therefore, an attempt to find an approximation to the sampling distribution of the Hill's estimator is required. To this end, a number of innovative techniques in the statistical analysis of extreme values proposes to adopt the powerful bootstrap tool to find the optimal number of order statistics adaptively [6–9]. By capitalizing on these recent advances in the extreme value statistics literature, in this paper we adopt a subsample semi-parametric bootstrap algorithm in order to make a reasonable and more automated selection of the extreme quantiles useful for studying the upper tail of income distributions and to end up at less ambiguous estimates of α. This methodology is described in Section 2 and its application to Australian and Italian income data [10,11] is given in Section 3. Some conclusive remarks are reported in Section 4.

2. Estimation technique for threshold selection

In this section, we consider the problem of finding the optimal threshold u_n^*—or equivalently the optimal number k^* of extreme sample values above that threshold—to be used for estimation of α. In order to achieve this task, we minimize the *MSE* of the Hill's estimator (Eq. (1)) for a series of thresholds $u_n = y_{[n-k]}$, and pick the u_n value at which the *MSE* attains its minimum as u_n^*. Given that different threshold series choices define different sets of possible observations to be included in the upper tail of a specific observed sample $\mathbf{y}_n = \{y_i; i = 1, 2, \ldots, n\}$, only the observations exceeding a certain threshold that are additionally distributed according to a Pareto's cumulative distribution function $PD_{\hat{\alpha}_n(k),u_n}(y)$ are included in the series. In order to check this condition, we perform for each threshold in the original sample a *Kolmogorov–Smirnov* (*K–S*) goodness-of-fit test for the null hypothesis $H_0 : \hat{F}_n(y) = PD_{\hat{\alpha}_n(k),u_n}(y)$ versus the general alternative of the form $H_1 : \hat{F}_n(y) \neq PD_{\hat{\alpha}_n(k),u_n}(y)$, where $\hat{F}_n(y)$ is the empirical distribution function, and $\hat{\alpha}_n(k)$ is a prior estimate for each threshold u_n of the Pareto's tail index obtained through the Hill's statistic. Following the methodology in [12], the formal steps in making a test of H_0 are as follows:

(a) Calculate the original *K–S* test statistic D by using the formula

$$D = \sup_{-\infty < y < \infty} |\hat{F}_n(y) - PD_{\hat{\alpha}_n(k),u_n}(y)|.$$

(b) Calculate the modified form T^* by using the formula

$$T^* = D\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right). \tag{2}$$

(c) Reject H_0 if T^* exceeds the cutoff level, z, for the chosen significance level.

To obtain an estimate of finite-sample bias and variance (and thus *MSE*) at each threshold coming from the null hypothesis H_0, a natural criterion is to use the *bootstrap* [13]. In its purest form, the bootstrap involves