A new Q-Q plot and its application to income data

Una variante del QQ-plot applicata ai dati sul reddito

Agostino Tarsitano
Dipartimento di Economia e Statistica - Università degli Studi della Calabria
agotar@unical.it

Summary. Si propone una variante del QQ-plot per accertare la capacità di una distribuzione di probabilità a rappresentare una data distributione di redditi. Lo stesso grafico può inoltre fornire, in parallelo, una misura della diseguaglianza presente nella distribuzione stessa

Keywords: order statistics, direct search estimates, graphical techniques, Burr 3 model

1. Introduction

A common practice to analyze income data is to fit a model to data and then compute all the inequality indices that appear useful as functions of the estimated shape parameters. In a sense, the choice of the probability distribution and the choice of the inequality index are independent. This paper develops a technique which derives simultaneously the model for the distribution of incomes and the measure of economic inequality.

2. A quantile plot method

The study of income distributions may receive a valuable stimulus by a graphical technique, similar in spirit to a QQ plot. Let $H(z)$ a nondegenerate distribution such that $E(Z)<\infty$.

Suppose that one individual has income X. If the value X were included in a random sample of size n: $\{Z_1, Z_2,\ldots, Z_n\}$ from H and the observations ranked in ascending order: $Z_{i:n}$ for $i=1,2,\ldots,n$ with $E(Z_{i:n})=m_{i:n}$ then our individual is prepared to occupy the position determined according to the rule: $||X-m_{r:n}|| \leq ||X-m_{i:n}||$ for $i=1,2,\ldots,n$ that is, the position r whose expected income, under the model H, is nearest to X according to a fixed metric.

Let $\{X_1,X_2,\ldots,X_n\}$ be a random sample of incomes from F, that is the "true", but unknown law of income distribution and consider the scatterplot with coordinates

$$\left\{\left(\frac{m_{i:n}-m_n}{\tau_n}\right),\left(\frac{X_{i:n}-\mu_n}{\mu_n}\right)\right\};\quad i=1,2,\ldots,n;\qquad \tau_n=\frac{\sum_{i=1}^{n}\left(m_{i:n}-m_n\right)^2}{n\left(m_{n:n}-m_n\right)};\qquad (1)$$

where $m_n=n^{-1}\sum w_{i:n}$ and $\mu_n=n^{-1}\sum X_{i:n}>0$. The rationale behind the plot is that if the plotting points lie near a straight line through the origin this provides an informal validation of H to model the observed data. Departure from H would be evidenced by increased curvature. If the linearity of the plot is satisfactory, then the least squares estimates of the slope

$$\boldsymbol{G}(H) = n^{-1}\sum_{i=1}^{n}\left[\frac{m_{i:n}-m_n}{m_{n:n}-m_n}\right]\frac{X_{i:n}}{\mu_n} \tag{2}$$

can measure the inequality in the $\{X_{i:n}\}$. It is easily seen that each choice of H generates a reasonable inequality measure. First, $G(H)$ is scale invariant and obeys the Pigou-Dalton principle. In fact, the effect of a neutral transfer of $d>0$ from income j to the income i, with $i<j$, is $d(m_{i:n}-m_{j:n})/(m_{n:n}-m_n)<0$. Second, since both $\{X_{i:n}\}$ and $\{m_{i:n}\}$ are nondecreasing, then $G(H) \geq 0$. Also, $m_{i:n} \leq m_{n:n}$ then $G(H) \leq 1$. The lower limit $G(H)=0$ can be interpreted as a tendency to the situation in which the $\{X_{i:n}\}$ are close together. The upper limit $G(H)=1$ is achieved as n diverges while only individual gets all the income. Third, the weights $\{w_{i:n}\}$ reveal the normative preferences to the measurement of income inequality implicit in any $G(H)$. Several measures can be obtained as special cases of (2).

$$\begin{array}{llll} w_{i:n}=i & Gini & w_{i:n}=1+S_{i-1} & Bonferroni \\ w_{i:n}=i^2 & Piesch-Giaccardi & w_{i:n}=S_{n-1}-S_{n-i} & De\ Vergottini \\ w_{i:n}=1-\left(\frac{n-i}{n}\right)^2 & Mehran & & \end{array} \qquad where\ S_i=\sum_{j=1}^{i} j^{-1},\ S_0=0 \tag{3}$$

In particular, the rectangular $[0,1]$ distribution leads to the Gini index, the unit exponential and the reflected unit exponential are the premise, respectively, of the De Vergottini and the Bonferroni index. The Piesch-Giaccardi index and the Mehran index can be derived by the generalized beta of the first type.

2. Large sample approximation

The proposed methodology can be difficulty to apply because explicit algebraic expressions for $\{w_{i:n}\}$ and τ_n are rarely available. To alleviate this problem somewhat, we can derive the weight function directly from the quantile function of the presumed model $H(z)$. To this end we further assume that Z is an absolutely continuous random variable with associated density function $g(z)$ continuous and strictly positive on the support of Z. This allows us to exploit a well known convergence result $w_{(n+1)t:n} \rightarrow H^{-1}(t)$. Since the size of the samples encountered in income distribution analysis is fairly large, this approximation should be satisfactory. Thus, the restricted class of indices is defined as follows

$$G(H)=n^{-1}\sum_{i=1}^{n}\left[\frac{H^{-1}\left(\frac{i}{n+1}\right)-w_n}{H^{-1}\left(\frac{n}{n+1}\right)-w_n}\right]\frac{X_{i:n}}{\mu_n}; \qquad w_n=n^{-1}\sum_{i=1}^{n}H^{-1}\left(\frac{i}{n+1}\right) \tag{4}$$

The message that our proposal conveys is that one may select the best H distribution out of a finite family of admissible distributions which yields the most linear QQ-plot. For instance, let us consider the Burr/3 distribution

$$H(z)=\left[1+z^{-\frac{1}{\beta}}\right]^{-\frac{1}{\gamma}} \qquad z>0, \qquad \gamma>0,\ 0<\beta\leq 1 \tag{5}$$

The expected value of the order statistics can be approximated by $W(t;\beta,\gamma)=[t^{-\gamma}-1)^{-\beta}$. The discrete version of $W(t;\beta,\gamma)$ determines a parametric family of indices belonging to *(1)*.

$$T(\beta,\gamma)=n^{-1}\sum_{i=1}^{n}\left\{\frac{\left[\left(\frac{n+1}{i}\right)^{\gamma}-1\right]^{-\beta}-w_n}{\left[\left(\frac{n+1}{n}\right)^{\gamma}-1\right]^{-\beta}-w_n}\right\}\frac{X_{i:n}}{\mu_n}\ ; \qquad w_n=n^{-1}\sum_{i=1}^{n}\left[\left(\frac{n+1}{i}\right)^{\gamma}-1\right]^{-\beta} \tag{6}$$

The indices of class (6) offers a flexible, and general tool for the measurement of income inequality in empirical and theoretical distributions. In fact, one may choose (β,γ) to conform to one's judgement about social welfare. Nevertheless, they have the practical inconvenient of relying on specific parametric assumptions. The analyst is required to specify, by means of additional constrains, assumptions and/or observed facts a credible value of (β,γ) to refine (6) and thus obtain a unique index of inequality.

3. Numerical results and discussion

To illustrate the use of the procedure, we employed the data on net disposable income coming from the survey on Italian household budgets for the years 1993,1995,1998, 2000 (positive incomes only). To create the Burr/3 QQ-plot, an estimates of the parameters of (6) is necessary. To this end, we determined the pair (β,γ) which maximizes the linearity of the plot, that is the parameter combination maximizing

$$\rho(\beta,\gamma)=n^{-1}\sum_{i=1}^{n}\left\{w_n-\left[\left(\frac{n+1}{i}\right)^{\gamma}-1\right]^{-\beta}\right\}\frac{X_{i:n}}{s_x s_z}, \qquad s_z=\sqrt{n^{-1}\left\{\left[\left(\frac{n+1}{i}\right)^{\gamma}-1\right]^{-\beta}-w_n\right\}^2} \tag{7}$$

where s_x is the standard deviation of the $\{X_{i:n}\}$. More specifically, the value of (β,γ) was obtained performing a direct search algorithm in two stages. The first stage consisted of a quasi-random (Faure type) sequence of 10^6 points in the region $A=\{(\beta,\gamma)|\ 0<\beta\leq 1,\ 0<\gamma\leq 10\}$. For the second stage we applied the downhill simplex method. Both stages require only function evaluations. The simplex algorithm was started from the point where the first search has found a minimum. The results of this approach are summarized in Figure 1 which includes the five inequality measures in (3) computed from microdata, the size of the sample, the maximum value of $\rho(\beta,\gamma)$, the estimates of the parameters, and the corresponding value of (7). Since the $\{X_{i:n}\}$ are highly correlated and heteroschedastic, the

usual properties of the correlation coefficient do not apply and Monte Carlo methods must be used to find the sample distribution of (β,γ). Research is continuing along these lines.

Figure 1: *QQ-plot for Italian household budgets*

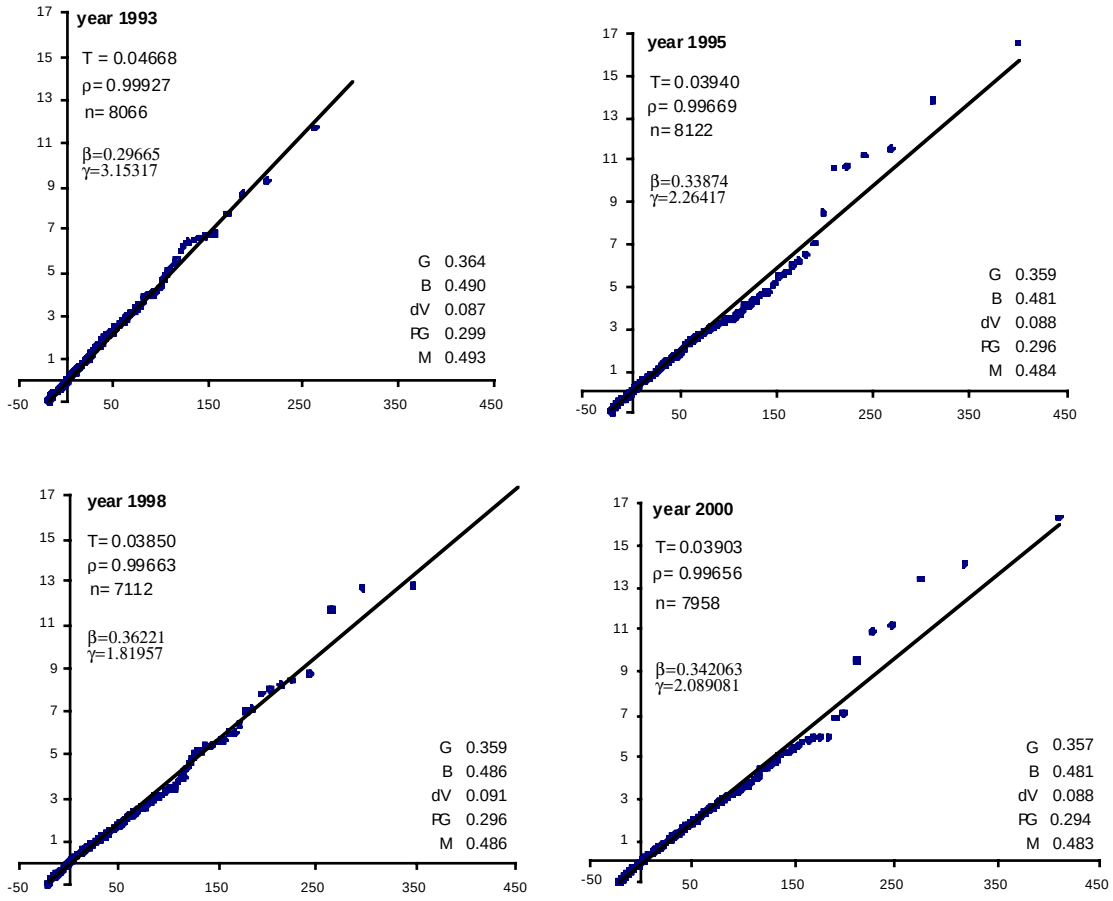

Figure 1 shows that the straight line provides a very good fit in all the QQ-plots with the exception of a small proportion of data in the tails which usually are very difficult to accommodate. It must be said that the QQ-plot is more sensitive to discrepancies in tail regions than in the middle of the hypothesized distribution; also, the value of (7) is stable across the years even though the estimate of the parameters undergoes to substantial variation. This reinforces the validity of the Burr/3 as a manageable model for describing a size distribution of incomes. The inequality decreases most between 1993 and 1995. In fact, the relative variation of $T(\beta,\gamma)$ is -15.6% which is very much higher than for the measures reported in (3). There is also unanimous agreement between these measures in indicating that inequality falls between 1995 and 1998, then rises between 1998 and 2000. Although the magnitude of the variation in $T(\beta,\gamma)$ is comparable with that of the other indices, its tendency is reversed thus telling an interestingly different story.

References

http://www.bancaditalia.it/statistiche