Nonparametric Kernel Estimation of the Size Distribution of Income

Tomson Ogwang
University of Northern British Columbia, Economics Program
3333 University Way
Prince George, BC, Canada V2N 4Z9
E-Mail Address: ogwang@unbc.ca

1. Introduction

Two important empirical questions commonly arise in the analyses of income distributions, namely the appearance of the density and the estimate of the size distribution of income. The Gini index is an important measure of the size distribution of income. In this paper a method of obtaining a single value estimate of the Gini index from ungrouped data is proposed. The proposed method is based on the Rosenblatt-Parzen kernel method of nonparametric probability density estimation (Rosenblatt, 1956; Parzen, 1962). Since the Gini index is defined in terms of the cumulative distribution function, the proposed method substitutes the kernel estimates of the distribution function for the definition of the Gini index.

2. Derivation of kernel estimates of the Gini index

To derive a general formula for kernel estimates of the Gini index, hereafter referred to as the Nonparametric Gini (NPG), it is assumed that $x_1, x_2, \ldots, x_T$ constitute a random sample of T observations from a particular strictly continuous distribution of income whose underlying density function, f(x), and, hence, the cumulative distribution function, F(x), are unknown. Hereafter, x_i is used to denote the income of the i-th income receiving unit (individual, household, etc.).

Following Silverman (1986, p. 15), the kernel estimator of f(x) is given by

$$\hat{f}(x) = (Th)^{-1} \sum_{i=1}^{T} k\left(\frac{x - x_i}{h}\right) \tag{1}$$

where k(.) is a symmetric, integrable weighting function called the kernel function and the quantity h is called the smoothing parameter or bandwidth.

The standard formula for the Gini index (for example, Lerman and Yitzhaki, 1984) is given by

$$G = \frac{1}{m} \int_a^b F(x)[1 - F(x)]dx \tag{2}$$

where a and b are the lowest and the highest incomes, respectively, and ： is the mean income.

Following Silverman (1986, p. 148) the kernel estimator of F(x) is given by

$$\hat{F}(x) = T^{-1} \sum_{i-1}^{T} K\left(\frac{x - x_i}{h}\right) \tag{3}$$

where K(.) is the cumulative distribution function of the kernel function.
The kernel estimator of ： is given by

$$\hat{m} = \sum_{i=1}^{T} x_i / T \tag{4}$$

Replacing F(x) and ： in equation (2) by their corresponding kernel estimators given by (3) and (4), respectively, we obtain the following expression for NPG

$$NPG = 1/\hat{m}\int_a^b \hat{F}(x)\left[1-\hat{F}(x)\right]dx = \left(\frac{T}{\sum_{i=1}^{T} x_i}\right)\int_a^b T^{-1}\sum_{i=1}^{T} K\left(\frac{x-x_i}{h}\right)\left[1-T^{-1}\sum_{j=1}^{T} K\left(\frac{x-x_j}{h}\right)\right]dx \quad (5)$$

Equation (5) can be rewritten as

$$NPG = \frac{h}{\sum_{i=1}^{T} x_i}\sum_{i=1}^{T}\int_{\frac{a-x_i}{h}}^{\frac{b-x_i}{h}} K(u)du - \frac{h}{T\sum_{i=1}^{T} x_i}\sum_{j=1}^{T}\int_{\frac{a-x_j}{h}}^{\frac{b-x_j}{h}} K^2(u)du - \frac{2}{T\sum_{i=1}^{T} x_i}\sum_{i}^{T}\sum_{<j}^{T}\int_a^b K\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_j}{h}\right)dx \quad (6)$$

A review of the literature on kernel estimation (e.g. Silverman, 1986) suggests that the choice of the kernel function is less crucial relative to the choice of the bandwidth parameter. It is well known (e.g. Azzalini, 1981) that the asymptotically optimal bandwidth parameter for estimating the distribution function is proportional to $T^{-1/3}$, where T is the sample size. In contrast, the asymptotically optimal bandwidth parameter for estimating a probability density function is proportional to $T^{-1/5}$ (see Parzen, 1962, Lemma 4A). In the following illustrative example and the design of the Monte Carlo experiment, the Epanechnikov kernel (Silverman, 1986, p. 43) is employed and the bandwidth is obtained using the formula $h = sT^{-1/3}$, where s is the standard deviation of the incomes computed from the sample.

3. Illustrative Example

In order to demonstrate the computation of NPG using equation (6), we utilized ungrouped data on the incomes of 10,938 households taken from the 1982 Canadian income data from the Family Expenditure Survey microdata tapes. The resulting kernel estimate of the Gini index was 0.344027. A preliminary assessment of the reliability of NPG was made by determining whether the estimates lie within Gastwirth's bounds (Gastwirth, 1972). The computed values of Gastwirth's lower and upper bounds, based on the 15 income classes, were 0.334707 and 0.348974, respectively, in which case our estimate of NPG is within the bounds.

4. Monte Carlo Experiment

We also undertook a Monte Carlo study to compare the performance of our estimator with those of the regression methods suggested by Lerman and Yitzhaki (1984). To this end, we generated 1000 samples of size 1000 and 5000 in accordance with a gamma density function

$$f(x;a,b) = \frac{\left(b^a x^{a-1} e^{-bx}\right)}{\Gamma(a)} \quad (a>0;\ b>0;\ x>0) \quad (11)$$

For an expression for the Gini index corresponding to the gamma density function, see McDonald and Ransom (1981). The gamma density function was chosen since its parameters can be changed to mimic several well-known distributions. For purposes of experimentation five different combinations of the parameters " and $ in equation (11) were used, namely: (i) "=1, $=1/2 (ii) "=1, $=5 (iii) "=2, $=5 (iv) "=10, $=5 and (v) "=20, $=5. It is noteworthy that the distribution becomes more symmetric as we move from cases (ii) to (v).

The Mean Squared Errors, reported in the table below, are smaller for the kernel estimator than for the regression estimator in cases (i), (ii) and (iii). However, the regression estimator outperforms the kernel estimator in cases (iv) and (v) where the distribution becomes more symmetric.

Mean squared errors of estimators for the Gini index (x 10^{-2})

	True Gini Index	Sample Size 1000		Sample Size 5000	
		Kernel	Regression	Kernel	Regression
(i) " =1, $=1/2	0.5	2.2466	2.6077	2.2377	2.5809
(ii) " =1, $=5	0.5	2.2466	2.6077	2.2377	2.5809
(iii) " =2, $=5	0.375	0.0904	0.14	0.0869	0.1286
(iv) " =10, $=5	0.1762	3.0465	2.7525	2.8818	2.6657
(v) " =20, $=5	0.1254	5.0749	4.5625	4.846	4.582

REFERENCES

Azzalini, A. (1981), "A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method," Biometrika, 68, 326-328.

Gastwirth, J.L. (1972), " The Estimation of the Lorenz Curve and Gini Index," Review of Economics and Statistics, 54, 306-316.

Lerman, R.I. and S. Yitzhaki, (1984), " A Note on the Calculation and Interpretation of the Gini Index," Economics Letters, 15, 363-368.

McDonald J.B. and M.R. Ransom (1981), "An Analysis of the Bounds for the Gini Coefficient", Journal of Econometrics, 17, 177-188.

Parzen, E. (1962), "On the Estimation of Probability Density and Mode," Annals of Mathematical Statistics, 27, 1065-1076.

Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of the Density Function," Annals of Mathematical Statistics, 27, 832-837.

Silverman, B.W. (1986), Density Estimation for Statistics and Data Analysis, New York, Chapman and Hall.

RESUME

In this paper a nonparametric smoothing procedure for evaluating the Gini index from individual observations is proposed. The method is applied to Canadian income data with satisfactory results. Furthermore, the results of Monte Carlo experiments conducted show a superior performance of the kernel estimator of the Gini index to the regression estimator in cases where the underlying distribution of income is highly skewed.