ON CALCULATION OF THE EXTENDED GINI COEFFICIENT

BY DUANGKAMON CHOTIKAPANICH

Curtin University of Technology

AND

WILLIAM GRIFFITHS

University of Melbourne

The conventional formula for estimating the extended Gini coefficient is a covariance formula provided by Lerman and Yitzhaki (1989). We suggest an alternative estimator, obtained by approximating the Lorenz curve by a series of linear segments. In a Monte Carlo experiment designed to assess the relative bias and efficiency of the two estimators, we find that, when using grouped data with 20 or fewer groups, our new estimator has less bias and lower mean squared error than the covariance estimator. When individual observations are used, or the number of groups is 30 or more, there is little or no difference in the performance of the two estimators.

1. INTRODUCTION

A generalization of the Gini coefficient, called the extended Gini coefficient, was introduced by Yitzhaki (1983) to accommodate differing aversions to inequality. While a number of algebraically-equivalent formulae have been described in the literature for estimating the original Gini coefficient (for example, Nygård and Sandström, 1981, Table 8.1; Creedy, 1996, pp. 10, 20), estimation of the extended Gini coefficient seems to have been confined to a covariance formula suggested by Lerman and Yitzhaki (1989). We suggest an alternative estimator, obtained by approximating the Lorenz curve by a series of linear segments. The covariance formula and our linear-segment estimator, are identical for the original Gini coefficient, but are not equal in general for the extended Gini coefficient. Thus, for the original Gini coefficient, any choice between the two estimators is made on the basis of computational convenience only. For the extended Gini coefficient, however, both computational convenience and estimator sampling properties are important considerations. In a Monte Carlo experiment that we conduct, the two estimators have similar properties when calculated from individual observations; when calculated from grouped data, our new estimator outperforms the covariance estimator in terms of both bias and mean-squared error. Our results have relevance not just for estimation of the extended Gini coefficient, but also for estimation of social welfare measures that are dependent on the extended Gini coefficient. See, for example, Lambert (1993, pp. 123–30).

In Section 2 we introduce required notation and describe two versions of the original Gini coefficient. In Section 3 we present the extended Gini coefficient and its corresponding covariance estimator, and go on to derive our alternative

estimator, leaving some of the details to an appendix. The set-ups and results of the Monte Carlo experiment are described in Section 4 and some summary remarks are made in Section 5.

2. The Gini Coefficient

Let $\pi = F(x)$ represent the distribution function for income x and let $\eta = F_1(x)$ be the corresponding first moment distribution function. The relationship between η and π, defined for $0 \leq x < \infty$ is the Lorenz curve. We denote it by $\eta = L(\pi)$. The much-used Gini coefficient is equal to twice the area between a 45-degree line and the Lorenz curve. That is,

$$(1) \qquad G = 1 - 2 \int_0^1 L(\pi)\, d\pi.$$

It can also be written as (see, for example, Lambert, 1993, p. 43):

$$(2) \qquad G = -1 + \frac{2}{\mu_x} \int_0^\infty x F(x) f(x)\, dx$$

$$= \frac{2}{\mu_x} \text{cov}\{x, F(x)\}$$

where $\mu_x = E(x)$ is mean income and $f(x) = dF(x)/dx$ is the density function for income.

Algebraically-equivalent discrete versions of equations (1) and (2) are often used to estimate G. To introduce the notation necessary to describe these two estimators, suppose that income data have been sampled and classified into M income groups. The estimators that we describe can be used with grouped data or with individual observations. In the case of individual observations, M is the number of observations, and, in what follows, there is one observation in each "group", with the proportion of observations in each group being $p_i = 1/M$. Given this level of generality, we assume the following information is available for the i-th group:
1. Average income x_i.
2. The proportion of observations p_i.
3. The cumulative proportion of observations $\pi_i = p_1 + p_2 + \cdots + p_i$.
4. The proportion of income $\phi_i = p_i x_i / \sum_{j=1}^M p_j x_j$.
5. The cumulative proportion of income $\eta_i = \phi_1 + \phi_2 + \cdots + \phi_i$.

Also, let $\bar{x} = \sum_{i=1}^M p_i x_i$ denote the sample mean income.

As noted by Lerman and Yitzhaki (1989), the discrete version of (2) that provides an estimator for G, is

$$(3) \qquad \hat{G}_1 = \frac{2}{\bar{x}} \sum_{i=1}^M p_i (x_i - \bar{x})(\hat{\pi}_i - \bar{\pi})$$

where $\hat{\pi}_i = (\pi_{i-1} + \pi_i)/2$ and $\bar{\pi} = \sum_{i=1}^M p_i \hat{\pi}_i$.

To obtain a discrete version of equation (1) to use as an estimator for G, the Lorenz curve $L(\pi)$ is approximated by a number of linear segments, with the i-th

linear segment being a straight line joining (π_{i-1}, η_{i-1}) to (π_i, η_i). Then, the area defined by equation (1) can be estimated by aggregating the areas between the linear segments and the 45-degree line. This process leads to another familiar expression for the Gini coefficient:

$$(4) \qquad \hat{G}_2 = \sum_{i=1}^{M-1} \eta_{i+1}\pi_i - \sum_{i=1}^{M-1} \eta_i\pi_{i+1}.$$

It can be shown that $\hat{G}_1 = \hat{G}_2$. However, when the estimation principles used to obtain $\hat{G}_1$ and $\hat{G}_2$ are applied to the extended Gini coefficient introduced by Yitzhaki (1983), they yield estimators that are, in general, not identical. Previous literature has focused on a covariance formula similar to $\hat{G}_1$ (Lerman and Yitzhaki, 1989). The purpose of our paper is to derive an expression for the extended Gini counterpart of $\hat{G}_2$ and to compare the bias and efficiency of the two alternative estimators via a Monte Carlo experiment.

3. A New Estimator for the Extended Gini Coefficient

The extended Gini coefficient can be written as

$$(5) \qquad G(v) = 1 - v(v-1) \int_0^1 (1-\pi)^{v-2} L(\pi)\, d\pi$$

$$= 1 - \frac{v}{\mu_x} \int_0^\infty x[1 - F(x)]^{v-1} f(x)\, dx$$

$$(6) \qquad = -\frac{v}{\mu_x} \operatorname{cov}\{x, [1 - F(x)]^{v-1}\}$$

where v is an inequality aversion parameter. The coefficient $G(v)$ is defined for $v > 1$ and is equal to the original Gini coefficient when $v = 2$.

The covariance-formula estimator, given by the empirical discrete version of equation (6) is (Lerman and Yitzhaki, 1989):

$$(7) \qquad \hat{G}_1(v) = -\frac{v}{\bar{x}} \sum_{i=1}^{M} p_i(x_i - \bar{x})[(1 - \hat{\pi}_i)^{v-1} - m]$$

where $m = \sum_{i=1}^{M} p_i(1 - \hat{\pi}_i)^{v-1}$.

To derive an alternative estimator obtained by approximating the Lorenz curve in equation (5) with a series of linear segments, we write the equation of a linear segment from (π_{i-1}, η_{i-1}) to (π_i, η_i) as $\eta = c_i\pi + d_i$ where $c_i = \phi_i/p_i$ and $d_i = (\pi_i\eta_{i-1} - \pi_{i-1}\eta_i)/p_i$. Then, a linear-segment approximation to $G(v)$ is given by:

$$(8) \qquad \hat{G}_2(v) = 1 - v(v-1) \sum_{i=1}^{M} \left(\int_{\pi_{i-1}}^{\pi_i} (1-\pi)^{v-2}(c_i\pi + d_i)\, d\pi \right).$$

In the appendix we show that this expression reduces to:

$$(9) \qquad \hat{G}_2(v) = 1 + \sum_{i=1}^{M} \left(\frac{\phi_i}{p_i}\right)[(1-\pi_i)^v - (1-\pi_{i-1})^v].$$

543

This expression is a relatively simple one which is easy to calculate, despite the tedious algebra necessary to derive it. Its sampling properties are assessed in Section 4. It can be shown that $\hat{G}_1(v) = \hat{G}_2(v)$ if $v = 2$. However, in general, the two estimators are not identical.

4. THE RELATIVE PERFORMANCE OF THE TWO ESTIMATORS

Given the existence of two reasonable alternative estimators for the extended Gini coefficient, their relative sampling performance is of interest. To evaluate this performance, we report the results of a Monte Carlo experiment with two hypothetical income distributions. One distribution is a lognormal distribution where $\log(x)$ is normally distributed with mean $\mu = 5$ and standard deviation $\sigma = 1.5$. The second distribution is one suggested by Singh and Maddala (1976), with distribution function

$$\pi = F(x) = 1 - \frac{1}{\left(1 + \left(\frac{x}{b}\right)^a\right)^q} \qquad a = 0.84, b = 400, q = 2.4.$$

Both these distributions exhibit a similar and relatively high level of inequality with, approximately, $G(1.33) = 0.43$, $G(2) = 0.71$ and $G(5) = 0.92$. Monte Carlo results were also obtained for other parameterizations, with lower levels of inequality. These results are available from the authors upon request. They lead to the same conclusions as the results reported here.

The other dimensions over which sensitivity was assessed were the value of v and the number of income groups. For v, we used $v = (1.33, 1.67, 2, 3, 5)$. Sampling performance was evaluated by drawing 5000 samples, each of size 2000, from each distribution. In addition to using the individual observations ($M = 2000$), results were obtained for three income groupings, $M = (10, 20, 30)$.

The results from the Monte Carlo experiment appear in Tables 1 and 2. The bias of the two estimators appears in Table 1. Their relative variance, and their relative mean-squared error appear in Table 2. Values of relative variance and mean-squared error greater than one imply the covariance estimator $\hat{G}_1(v)$ is out-performing our linear-segment estimator $\hat{G}_2(v)$.

From Table 1 we can make the following observations about bias:
1. The bias of both estimators is always negative, reflecting the fact they implicitly assume no inequality within each group.
2. When $M = 2000$, both estimators have negligible and almost identical bias; the bias is also relatively small for $M = 30$.
3. The absolute bias of the covariance estimator is never less, and often substantially more, than the absolute bias of the linear-segment estimator.
4. The relative performance of the linear-segment estimator improves the further is the departure of v from 2, and the smaller the number of groups M.

From the results in Table 2, we see that the lower bias for the linear-segment estimator comes at a cost of higher variance. Since a comparison of biases favors the linear-segment estimator, and a comparison of variances favors the covariance

TABLE 1
Bias of the Estimators

| Groups | Estimator | v | | | | |
		1.33	1.67	2	3	5
Lognormal						
M = 10	$\hat{G}_1(v)$	−0.020	−0.013	−0.010	−0.012	−0.021
	$\hat{G}_2(v)$	−0.016	−0.012	−0.010	−0.008	−0.007
M = 20	$\hat{G}_1(v)$	−0.009	−0.005	−0.003	−0.004	−0.007
	$\hat{G}_2(v)$	−0.007	−0.004	−0.003	−0.003	−0.002
M = 30	$\hat{G}_1(v)$	−0.006	−0.003	−0.002	−0.002	−0.003
	$\hat{G}_2(v)$	−0.004	−0.003	−0.002	−0.002	−0.001
M = 2000	$\hat{G}_1(v)$	−0.002	−0.001	−0.001	−0.001	−0.000
	$\hat{G}_2(v)$	−0.002	−0.001	−0.001	−0.001	−0.000
Singh–Maddala						
M = 10	$\hat{G}_1(v)$	−0.024	−0.015	−0.011	−0.013	−0.022
	$\hat{G}_2(v)$	−0.020	−0.014	−0.011	−0.009	−0.008
M = 20	$\hat{G}_1(v)$	−0.012	−0.007	−0.005	−0.005	−0.007
	$\hat{G}_2(v)$	−0.010	−0.006	−0.005	−0.004	−0.003
M = 30	$\hat{G}_1(v)$	−0.008	−0.005	−0.004	−0.003	−0.004
	$\hat{G}_2(v)$	−0.007	−0.005	−0.004	−0.002	−0.002
M = 2000	$\hat{G}_1(v)$	−0.005	−0.003	−0.002	−0.001	−0.001
	$\hat{G}_2(v)$	−0.004	−0.003	−0.002	−0.001	−0.001

TABLE 2
Relative Variance $\mathrm{var}\,[\hat{G}_2(v)]/\mathrm{var}\,[\hat{G}_1(v)]$ and Relative Mean Squared Error $\mathrm{MSE}\,[\hat{G}_2(v)]/\mathrm{MSE}\,[\hat{G}_1(v)]$

| Groups | v | | | | |
	1.33	1.67	2	3	5
Lognormal					
M = 10	1.038	1.003	1.000	1.024	1.087
	(0.815)	(0.964)	(1.000)	(0.715)	(0.178)
M = 20	1.031	1.003	1.000	1.007	1.025
	(0.955)	(0.996)	(1.000)	(0.940)	(0.494)
M = 30	1.021	1.002	1.000	1.003	1.012
	(0.993)	(0.999)	(1.000)	(0.983)	(0.791)
M = 2000	1.008	1.001	1.000	1.000	1.000
	(1.006)	(1.000)	(1.000)	(1.000)	(1.000)
Singh–Maddala					
M = 10	1.039	1.003	1.000	1.026	1.097
	(0.874)	(0.975)	(1.000)	(0.772)	(0.204)
M = 20	1.035	1.004	1.000	1.007	1.029
	(0.978)	(0.998)	(1.000)	(0.954)	(0.537)
M = 30	1.023	1.002	1.000	1.003	1.013
	(1.003)	(1.000)	(1.000)	(0.986)	(0.814)
M = 2000	1.012	1.001	1.000	1.000	1.000
	(1.010)	(1.001)	(1.000)	(1.000)	(1.000)

Note: The relative MSEs appear in parentheses below the relative variances.

estimator, a mean-squared error comparison is useful. The results using this criterion appear in parentheses in Table 2. These results show that:

1. For $M = 30$ and $M = 2000$ the performance of the two estimators is very similar except when $v = 5$ and $M = 30$, where the linear-segment estimator is noticeably better.
2. For $M = 10$ and $M = 20$ the linear segment estimator is always better, and sometimes very much better than the covariance estimator.

5. Summary

An estimator for the extended Gini coefficient has been derived by approximating the Lorenz curve by a series of linear segments. This estimator is simple to compute and has less bias than a covariance-based estimator that has been used in the literature. For grouped data where the number of groups is 20 or less, it also has lower mean-squared error than the covariance estimator. The experimental evidence is sufficiently strong to recommend that, for grouped data where the number of groups is 20 or less, practitioners should use our new estimator in preference to the covariance estimator. If the number of groups is 30 or more, or individual observations are available, both estimators perform equally well. Finally, it should be emphasized that both estimators require knowledge of arithmetic mean income in each group. For situations where these values are not available, the effect of using an alternative, such as the class midpoints, still needs to be investigated.

Appendix

In this appendix we show that equation (8) can be simplified to equation (9). The summation in equation (8) can be written as:

$$\sum_{i=1}^{M} \left(\int_{\pi_{i-1}}^{\pi_i} (1 - \pi)^{v-2}(c_i \pi + d_i) \, d\pi \right) = \sum_{i=1}^{M} [I_1(i) + I_2(i)]$$

where:

$$I_1(i) = c_i \int_{\pi_{i-1}}^{\pi_i} \pi(1-\pi)^{v-2} \, d\pi$$

$$= \frac{-c_i}{v-1} [\pi_i(1-\pi_i)^{v-1} - \pi_{i-1}(1-\pi_{i-1})^{v-1}] - \frac{c_i}{v(v-1)} [(1-\pi_i)^v - (1-\pi_{i-1})^v]$$

$$I_2(i) = d_i \int_{\pi_{i-1}}^{\pi_i} (1-\pi)^{v-2} \, d\pi = -\frac{d_i}{v-1} [(1-\pi_i)^{v-1} + (1-\pi_{i-1})^{v-1}].$$

Substituting for c_i and d_i, and adding these two equations, yields, after some algebra,

$$I_1(i) + I_2(i) = -\frac{1}{v-1}\left[\eta_i(1-\pi_i)^{v-1} - \eta_{i-1}(1-\pi_{i-1})^{v-1}\right]$$

$$-\frac{1}{v(v-1)}\left(\frac{\phi_i}{p_i}\right)[(1-\pi_i)^v - (1-\pi_{i-1})^v].$$

Summing this expression over all groups, we obtain:

$$\sum_{i=1}^{M}[I_1(i) + I_2(i)] = -\frac{1}{v(v-1)}\sum_{i=1}^{M}\left(\frac{\phi_i}{p_i}\right)[(1-\pi_i)^v - (1-\pi_{i-1})^{v-1}].$$

Substituting this expression into equation (8) gives the desired result.

REFERENCES

Creedy, J., *Fiscal Policy and Social Welfare*, Edward Elgar, Cheltenham, 1996.

Lambert, P. J., *The Distribution and Redistribution of Income: A Mathematical Analysis*, 2nd edition, Manchester University Press, Manchester, 1993.

Lerman, R. I. and S. Yitzhaki, "Improving the Accuracy of Estimates of Gini Coefficients," *Journal of Econometrics*, 42(1), 43–47, September, 1989.

Nygård, F. and A. Sandström, *Measuring Income Inequality*, Almqvist and Wiksell, Stockholm, 1981.

Singh, S. K. and G. S. Maddala, "A Function for Size Distribution of Incomes," *Econometrica*, 44(5), 963–70, September, 1976.

Yitzhaki, S., "On an Extension of the Gini Inequality Index," *International Economic Review*, 24(3), 617–28, October, 1983.