University of New England

School of Economic Studies

ON CALCULATION OF THE EXTENDED GINI COEFFICIENT

by

Duangkamon Chotikapanich and William Griffiths

No. 2000-1 – May 2000

Working Paper Series in Econometrics and Applied Statistics

ISSN 0157-0188

http://www.une.edu.au/febl/EconStud/wps.htm

 ISBN 1 86389 681 3

ON CALCULATION OF THE EXTENDED GINI COEFFICIENT

Duangkamon Chotikapanich and William Griffiths**

Abstract

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. The two estimators are identical for the original Gini coefficient, where the inequality aversion parameter is $v = 2$. However, for $v \neq 2$, they are different. In a Monte Carlo experiment designed to assess the relative bias and efficiency of the two estimators, we find that, when using grouped data, our new estimator has less bias and lower mean squared error than the covariance estimator. When single observations are used, there is little or no difference in the performance of the two estimators.

Key Words: Lorenz curve, income inequality

** Duangkamon Chotikapanich is a Senior Lecturer in the School of Economics and Finance at Curtin University of Technology. William Griffiths is a Professor in the School of Economic Studies at the University of New England. Contact information: School of Economic Studies, University of New England, Armidale, NSW 2351, Australia. Email: wgriffit@metz.une.edu.au.

ON CALCULATION OF THE EXTENDED GINI COEFFICIENT

1. Motivation

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 $0 \le x < \infty$ is the Lorenz curve. We denote it by $\eta = L(\pi)$. A much-used measure of income inequality is the Gini coefficient which is equal to twice the area between a 45-degree line and the Lorenz curve. That is,

$$G = 1 - 2\int_0^1 L(\pi)d\pi \tag{1}$$

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

$$\begin{aligned} G &= -1 + \frac{2}{\mu_x}\int_0^\infty xF(x)f(x)\,dx \\ &= \frac{2}{\mu_x}\operatorname{cov}\{x, F(x)\} \end{aligned} \tag{2}$$

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

Given a sample of observations on income, which may be available as single observations, or grouped into income classes, several algebraically equivalent formulas for estimating G have been described in the literature. See, for example, Table 8.1 in Nygård and Sandström (1981) or Creedy (1996, p.10, 20). The focus in this paper is on discrete versions of the expressions in equations (1) and (2), and on generalisations of them that can be used to estimate the extended Gini coefficient introduced by Yitzhaki (1983) to accommodate differing aversions to inequality. Because the discrete versions of (1) and (2) are algebraically equivalent, any choice between them is made simply on the basis of computational convenience. However, as we will see, for estimating the *extended* Gini coefficient, generalisations of the discrete formulas lead to two different estimators which are not algebraically equivalent. In these circumstances, when choosing a formula for estimation, computational convenience *and* estimator sampling properties are important considerations. A generalisation of the discrete version of the covariance formula in

(2) was suggested by Lerman and Yitzhaki (1989). In this paper we derive an alternative estimator which is based on a generalisation of a discrete version of equation (1). This alternative estimator is simple to calculate and has good sampling properties. In a Monte Carlo experiment that we conduct, the two estimators have similar properties when calculated from single 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, p.123-130).

In the remainder of this section we introduce required notation and give the discrete versions of (1) and (2) that have been used for estimating G. In Section 2 we present the extended Gini coefficient and its corresponding covariance estimator, and go on to derive our alternative estimator. The setups and results of the Monte Carlo experiment are described in Section 3 and some summary remarks are made in Section 4.

To introduce the notation necessary to describe the various 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 single observations. In the case of single ungrouped 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 $\pi_i = p_1 + p_2 + \ldots + p_i$.
4. The proportion of income $\phi_i = p_i x_i \Big/ \sum_{j=1}^{M} p_j x_j$.
5. The cumulative proportion of income $\eta_i = \phi_1 + \phi_2 + \ldots + \phi_i$.

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

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

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

where

$$\hat{\pi}_i = (\pi_{i-1} + \pi_i)/2$$

and

$$\bar{\pi} = \sum_{i=1}^{M} p_i \hat{\pi}_i$$

Another way to approach the estimation problem is via equation (1). If 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

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

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 note 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.

2. A New Estimator for the Extended Gini Coefficient

The extended Gini coefficient, proposed by Yitzhaki (1983) to accommodate the fact that different individuals can have different aversions to inequality, can be written as

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

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

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

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)

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

where

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

In the remainder of this section, we derive an alternative estimator obtained by approximating the Lorenz curve in equation (5) with a series of linear segments. To begin, note that the equation of a straight line joining (π_{i-1}, η_{i-1}) to (π_i, η_i) can be written as

$$\eta = c_i \pi + d_i \tag{8}$$

where

$$c_i = \frac{\eta_i - \eta_{i-1}}{\pi_i - \pi_{i-1}} = \frac{\phi_i}{p_i} \tag{9}$$

$$d_i = \frac{\pi_i \eta_{i-1} - \pi_{i-1}\eta_i}{\pi_i - \pi_{i-1}} \tag{10}$$

Now, denote the integral in equation (5) by $H(v)$. That is,

$$H(v) = \int_0^1 (1-\pi)^{v-2} L(\pi) d\pi \tag{11}$$

The linear-segment approximation to $H(v)$ is

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

$$= \sum_{i=1}^{M} [I_1(i) + I_2(i)] \tag{12}$$

where

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

$$= \frac{-c_i}{v-1}[\pi_i(1-\pi_i)^{v-1} - \pi_{i-1}(1-\pi_{i-1})^{v-1}] - \frac{c_i}{v(v-1)}[(1-\pi_i)^v - (1-\pi_{i-1})^v] \tag{13}$$

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

$$= -\frac{d_i}{v-1}[(1-\pi_i)^{v-1} + (1-\pi_{i-1})^{v-1}] \tag{14}$$

Substituting for c_i and d_i in equations (13) and (14), and adding these two equations, yields, after some algebra,

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

$$-\frac{1}{v(v-1)}\left(\frac{\phi_i}{p_i}\right)[(1-\pi_i)^v - (1-\pi_{i-1})^v] \tag{15}$$

Thus,

$$\hat{H}(v) = \sum_{i=1}^{M} [I_1(i) + I_2(i)]$$

$$= -\frac{1}{v(v-1)} \sum_{i=1}^{M} \left(\frac{\phi_i}{p_i}\right)[(1-\pi_i)^v - (1-\pi_{i-1})^{v-1}] \tag{16}$$

and the alternative estimator for the extended Gini coefficient is

$$\hat{G}_2(v) = 1 - v(v-1)\hat{H}(v)$$

$$= 1 + \sum_{i=1}^{M} \left(\frac{\phi_i}{p_i} \right) [(1-\pi_i)^v - (1-\pi_{i-1})^v] \quad (17)$$

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.

3. The Relative Performance of the Two Estimators

Given the existence of two very reasonable alternative estimators for the extended Gini coefficient, their relative sampling performance is of interest. To evaluate this performance, we set up a Monte Carlo experiment with 4 hypothetical income distributions. The details of those distributions appear in Table 1. Two are lognormal distributions where log(x) is normally distributed with mean μ and standard deviation σ. The other two are parameterisations of the distribution suggested by Singh and Maddala (1976); its distribution function is

$$\pi = F(x) = 1 - \frac{1}{\left(1 + \left(\frac{x}{b}\right)^a\right)^q}$$

The parameter values were chosen to give one lognormal and one Singh-Maddala distribution with relatively high inequality, and another pair of distributions with relatively low inequality. Setups 1 and 2 have relatively high inequality with approximate Gini coefficients of $G(2) = 0.71$. Setups 3 and 4 have relatively low inequality with approximate Gini coefficients of $G(2) = 0.38$. Relative to the lognormal distribution with a similar value of the Gini coefficient, the Singh-Maddala distribution has a thicker tail, with extreme values of income more likely. Designing the experiment in this way gives information on the sensitivity of performance to the type of income distribution and the level of inequality.

The other dimensions over which sensitivity was assessed were the value of v and the number of income groups. The chosen values of v are v = (1.33, 1.67, 2, 3, 5, 10). Values of the extended Gini coefficient corresponding to these values were computed via numerical integration of equation (6), and are reported in Table 1. Sampling performance was evaluated by drawing 5000 samples, each of size 2000, from each of the four distributions. In addition to using the single observations $(M = 2000)$, results were obtained for two income groupings $M = (10, 20)$.

The results from the Monte Carlo experiment appear in Tables 2, 3 and 4 which contain, respectively, the bias of the two estimators, their relative variance, and their relative mean-squared error. Values of relative variance and mean-squared error greater than one imply the covariance estimator $\hat{G}_1(v)$ is outperforming our linear-segment estimator $\hat{G}_2(v)$.

From Table 2 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.
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 3, we see that the lower bias for the linear-segment estimator usually comes at a cost of higher variance. Exceptions occur with the first three setups when $v = 10$ and $M = 10$. Since these exceptions look atypical relative to the remainder of the table, we carried out further experiments for values of v between 5 and 10 and greater than 10. These experiments revealed that, as v increases, there is a value for v beyond which the variance of the linear-segment estimator is less than the variance of the covariance estimator. The value depends on the setup; in setup 2 it is lower than for setups 1, 3 and 4. Also, it is larger for $M = 20$ than for $M = 10$. However, for $M = 10$ or 20, and for v large enough, there is always a reversal in the relative magnitudes of the variances. When $M = 2000$, there is no noticeable difference.

Since a comparison of biases favors the linear-segment estimator, and a comparison of variances favors the covariance estimator (except for large v), a mean-squared error comparison is useful. Using this criterion, the results in Table 4 show the performance of the linear-segment estimator is seldom worse, and sometimes very much better, than the covariance estimator. In the instances where the linear-segment estimator is worse (for example, $v = 1.33$, $M = 2000$, setup 2), it is only marginally worse, with the least favorable comparison being a mean-squared error which is only 0.8% larger.

4. 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 and lower mean-squared error than a covariance-based estimator that has been used in the literature. The experimental evidence is sufficiently strong to recommend that practitioners use our new estimator in preference to the covariance estimator in future empirical studies.

References

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

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

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

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

Singh, S.K. and G.S. Maddala (1976), "A Function for Size Distribution of Incomes", *Econometrica* 44, 963-970.

Yitzhaki, S. (1983), "On an Extension of the Gini Inequality Index", *International Economic Review* 24, 617-628.

Table 1. Setups for Monte Carlo Experiment

	Setup			
	1	2	3	4
Distribution	Lognormal	Singh Maddala	Lognormal	Singh Maddala
Parameters	$\mu = 5$ $\sigma = 1.5$	$b = 400$ $a = 0.84$ $q = 2.4$	$\mu = 6.4$ $\sigma = 0.7$	$b = 550$ $a = 2.9$ $q = 0.85$
μ_x	457.1	323.2	768.9	773.5
$G(1.33)$	0.4298	0.4360	0.1911	0.2034
$G(1.67)$	0.6160	0.6192	0.3057	0.3118
$G(2)$	0.7112	0.7143	0.3794	0.3785
$G(3)$	0.8360	0.8425	0.5059	0.4915
$G(5)$	0.9101	0.9211	0.6149	0.5919
$G(10)$	0.9543	0.9674	0.7112	0.6882

Table 2. Bias of the Estimators

Groups	Estimator	v					
		1.33	1.67	2	3	5	10
				Setup 1			
M = 10	$\hat{G}_1(v)$	-0.043	-0.025	-0.014	-0.007	-0.011	-0.041
	$\hat{G}_2(v)$	-0.035	-0.023	-0.014	-0.006	-0.003	-0.004
M = 20	$\hat{G}_1(v)$	-0.042	-0.023	-0.011	-0.003	-0.003	-0.007
	$\hat{G}_2(v)$	-0.033	-0.021	-0.011	-0.003	-0.001	-0.001
M = 2000	$\hat{G}_1(v)$	-0.002	-0.001	-0.001	-0.001	-0.000	-0.000
	$\hat{G}_2(v)$	-0.002	-0.001	-0.001	-0.001	-0.000	-0.000
				Setup 2			
M = 10	$\hat{G}_1(v)$	-0.035	-0.020	-0.012	-0.008	-0.020	-0.098
	$\hat{G}_2(v)$	-0.029	-0.018	-0.012	-0.006	-0.005	-0.009
M = 20	$\hat{G}_1(v)$	-0.033	-0.017	-0.008	-0.003	-0.004	-0.020
	$\hat{G}_2(v)$	-0.027	-0.015	-0.008	-0.003	-0.002	-0.002
M = 2000	$\hat{G}_1(v)$	-0.005	-0.003	-0.002	-0.001	-0.001	-0.000
	$\hat{G}_2(v)$	-0.005	-0.003	-0.002	-0.001	-0.001	-0.000
				Setup 3			
M = 10	$\hat{G}_1(v)$	-0.010	-0.007	-0.006	-0.007	-0.021	-0.086
	$\hat{G}_2(v)$	-0.007	-0.007	-0.006	-0.007	-0.012	-0.030
M = 20	$\hat{G}_1(v)$	-0.008	-0.005	-0.003	-0.002	-0.003	-0.013
	$\hat{G}_2(v)$	-0.008	-0.004	-0.003	-0.002	-0.002	-0.005
M = 2000	$\hat{G}_1(v)$	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	$\hat{G}_2(v)$	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
				Setup 4			
M = 10	$\hat{G}_1(v)$	-0.015	-0.009	-0.006	-0.007	-0.017	-0.062
	$\hat{G}_2(v)$	-0.012	-0.008	-0.006	-0.006	-0.010	-0.024
M = 20	$\hat{G}_1(v)$	-0.013	-0.007	-0.004	-0.002	-0.003	-0.010
	$\hat{G}_2(v)$	-0.010	-0.006	-0.004	-0.002	-0.002	-0.004
M = 2000	$\hat{G}_1(v)$	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	$\hat{G}_2(v)$	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001

Table 3. Relative Variance: $\mathrm{var}[\hat{G}_2(v)]/\mathrm{var}[\hat{G}_1(v)]$

	v					
Groups	1.33	1.67	2	3	5	10
			Setup 1			
M = 10	1.091	1.020	1.000	1.015	1.053	0.883
M = 20	1.096	1.023	1.000	1.004	1.010	1.031
M = 2000	1.008	1.000	1.000	1.000	1.000	1.000
			Setup 2			
M = 10	1.066	1.009	1.000	1.022	1.132	0.191
M = 20	1.072	1.013	1.000	1.007	1.033	1.133
M = 2000	1.011	1.001	1.000	1.000	1.000	1.000
			Setup 3			
M = 10	1.103	1.021	1.000	1.014	1.095	0.751
M = 20	1.119	1.027	1.000	1.000	1.015	1.153
M = 2000	1.001	1.000	1.000	1.000	1.000	1.000
			Setup 4			
M = 10	1.086	1.018	1.000	1.007	1.042	1.160
M = 20	1.092	1.021	1.000	1.000	1.006	1.049
M = 2000	1.006	1.000	1.000	1.000	1.000	1.000

Table 4. Relative MSE: $\text{MSE}[\hat{G}_2(v)] \big/ \text{MSE}[\hat{G}_1(v)]$

Groups	*v* 1.33	1.67	2	3	5	10
			Setup 1			
$M = 10$	0.685	0.863	1.000	0.917	0.288	0.017
$M = 20$	0.674	0.852	1.000	1.008	0.849	0.183
$M = 2000$	1.006	1.000	1.000	1.000	1.000	1.000
			Setup 2			
$M = 10$	0.773	0.933	1.000	0.883	0.167	0.009
$M = 20$	0.770	0.935	1.000	0.998	0.745	0.032
$M = 2000$	1.008	1.000	1.000	1.000	1.000	1.000
			Setup 3			
$M = 10$	0.581	0.855	1.000	0.889	0.398	0.122
$M = 20$	0.568	0.874	1.000	1.001	0.884	0.319
$M = 2000$	1.000	1.000	1.000	1.000	1.000	1.000
			Setup 4			
$M = 10$	0.731	0.931	1.000	0.950	0.551	0.172
$M = 20$	0.735	0.946	1.000	1.000	0.959	0.589
$M = 2000$	1.005	1.000	1.000	1.000	1.000	1.000