Journal of Statistical and Econometric Methods, vol.1, no.2, 2012, 31-38
ISSN: 2241-0384 (print), 2241-0376 (online)
Scienpress Ltd, 2012

Estimation of Gini coefficients using Lorenz curves

Johan Fellman[1,2]

Abstract

Primary income data yields the most exact estimates of the Gini coefficient. Using Lorenz curves, the Gini coefficient is defined as the ratio of the area between the diagonal and the Lorenz curve and the area of the whole triangle under the diagonal. Various attempts have been made to obtain accurate estimates. The trapezium rule is simple, but yields a positive bias for the area under the Lorenz curve and, consequently, a negative bias for the Gini coefficient. Simpson´s rule is better fitted to the Lorenz curve, but this rule demands an even number of subintervals of the same length. Lagrange polynomials of second degree can be considered as a generalisation of Simpson´s rule because they do not demand equidistant points. If the subintervals are of the same length, the Lagrange polynomial method is identical with Simpson´s rule. In this study, we compare different methods. When we apply Simpson´s rule, we mainly consider Lorenz curves with deciles. In addition, we use the trapezium rule, Lagrange polynomials and generalizations of Golden´s method (2008). No method is uniformly optimal, but the trapezium rule is almost always inferior and Simpson´s rule is superior. Golden´s method is usually of medium quality.

Mathematics Subject Classification: 62P20, 91B15, 91B82

---

[1] Swedish School of Economics and Business Administration, POB 479, FIN-00101 Helsinki, Finland e-mail: fellman@hanken.fi
[2] Folkhälsan Institute of Genetics, Department of Genetic Epidemiology, Helsinki, Finland

Article Info: *Received* : May 4, 2012. *Revised* : June 7, 2012
Published online : July 30, 2012

Keywords: Golden´s method, Lagrange polynomials, Pareto distribution, Simpson´s rule, Trapezium rule

1 Introduction

Primary income data yields the most accurate estimates of the Gini coefficient. However, the estimation must often be based on tables with grouped data or on Lorenz curves. The Lorenz curves are usually defined for five quintiles or for 10 deciles. If one uses the Lorenz curve, the Gini coefficient is defined as the ratio of the area between the diagonal and the Lorenz curve and the area of the whole triangle under the diagonal. For five quintiles, the trapezium rule is the most commonly used method. However, this rule yields positive bias for the estimate of the area under the Lorenz curve for every trapezium and, consequently, the rule causes negative bias for the Gini coefficient. Simpson´s rule is better fitted to the Lorenz curve, but demands an even number of subintervals of the same length. This means, for example, that Lorenz curves with 10 deciles are suitable. One has three L values for each doubled subinterval. The area under this part of the Lorenz curve is estimated so that the Lorenz curve is approximated by a parabola obtaining the same L values. Simpson´s rule obviously yields exact results for quadratic curves but, in general, this also holds for cubic curves. Lagrange polynomials of the second degree can be considered as a generalisation of Simpson´s rule and do not demand subintervals of equal length, but the number of subintervals should still be even. The polynomials obtained have to be integrated in order to yield approximate areas and Gini coefficients. If the subintervals are of the same length, the Lagrange polynomial method is identical with Simpson´s rule.

Various attempts have been made to produce more exact estimates. Gastwirth (1972) introduced interval estimates of the Gini coefficient in order to measure the accuracy of the estimates. Needleman´s study (1978) starts from the trapezium estimate of the Gini coefficient G_L. He then introduces an improved upper estimate G_U. His final estimate follows the "two-thirds rule" that is $G = \frac{G_L}{3} + \frac{2G_U}{3}$.

McDonald and Ransom (1981) considered the Γ density, applied Monte Carlo methods and introduced lower and upper bounds of the Gini estimates.

Golden (2008) showed how a quick approximation of the Gini coefficient can be calculated empirically, using numerical data in cumulative income quintiles. In this study, we intend to compare different methods. When we apply Simpson´s rule, we consider Lorenz curves with deciles. In addition, we use Lagrange polynomials and generalizations of Golden´s method.

2 Methods

There are several different situations and, consequently, alternative analyses of Gini coefficients have to be performed. When Lorenz curves are considered, the simplest situation is that they are defined for five quintiles or for 10 deciles. In the first case, the most commonly used method is the trapezium rule. For Simpson´s rule, the number of subintervals should be even and the intervals should have the same length. Consequently, the comparison of different rules can be performed for Lorenz curves with deciles.

Assume a Lorenz curve $L(p)$ with deciles. Let the observed values of the cumulative Lorenz curve be p_i and L_i for $i = 0, 1, ..., 10$. Note that $p_i = i/10$, ($i = 0, 1, ..., 10$), that $L_0 = 0$ and that $L_{10} = 1$. According to the trapezium rule, the estimated area under the Lorenz curve is

$$\tilde{I} = \tfrac{1}{2}\sum_{i=0}^{9}(L_{i+1} + L_i)(p_{i+1} - p_i) \qquad (1)$$

and the estimated Gini coefficient, G_T is $1 - 2\tilde{I}$. Every trapezium yields a positive bias to the estimated area, as can be seen in Figure 1. Since the biases obtained add and no elimination of biases can be performed, the estimated Gini coefficient always has a negative bias.

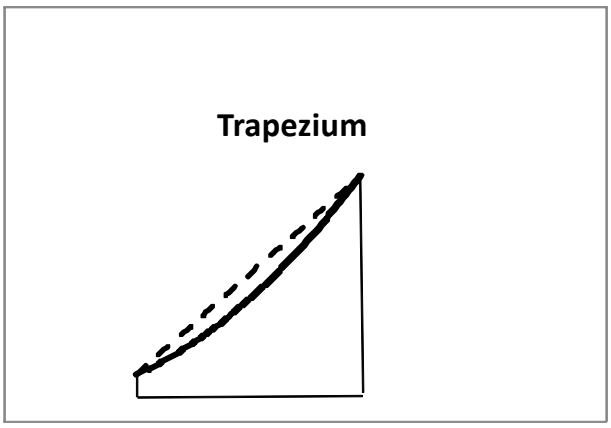

Figure 1: A sketch showing the bias in the trapezium rule

Compared to the trapezium rule, Simpson´s rule gives more accurate approximations. As stressed above, Simpson´s rule demands two restrictions: the number of subintervals has to be even and the subintervals have to be of equal length. In order to obtain Simpson´s rule, the subintervals should be grouped two by two. Each doubled subinterval has three *L* values. The area under this part of the Lorenz curve is estimated such that a parabola obtaining the same *L* values approximates the Lorenz curve. Assuming *2n* subintervals, the approximate area

formula for a doubled interval is $\tilde{I}_i = \frac{1}{3n}\left(L_i + 4L_{i+1} + L_{i+2}\right)$, the total sum is

$$\tilde{I} = \frac{1}{3n}\sum_{i=0}^{4}\left(L_{2i} + 4L_{2i+1} + L_{2i+2}\right) \qquad (2)$$

and $G_S = 1 - 2\tilde{I}$.

Golden (2008) gave a detailed account of an alternative method based on Lorenz curves with quintiles. He considered p and L in percentages. The layout of the method is presented in Table 1. First he determined where the cumulative income shortfall is greatest and defined Z as the largest quintile point of the cumulative income shortfall from perfect equality divided by 100. In order to obtain the largest cumulative income shortfall he defined the transformed variable $\tilde{L}_i = L_{i-1} + 20$. This transformation, $\tilde{L}_i = L_{i-1} + 20$, indicates a search for an interval at which L_i shifts from increases faster than p_i to slower increases. For low i´s, the transformed value $\tilde{L}_i > L_i$. Later, there is a first i value such that $\tilde{L}_i < L_i$. For this value, one finds an interval for which L is closely parallel with the diagonal, the greatest shortfall is obtained, and one defines $q = (20i - \tilde{L}_i)/100$. The estimated Gini coefficient in percentages, G_G, is $G_G = 50q(3-q)$. When this method was applied to 621 income observations, Golden (2008) noted that his approach performed better than the trapezium rule, also stressing that his method could be applied to Lorenz curves with deciles. Following Golden (1980), the data is given in percentages. The transformed variable $\tilde{L}_{20i}$ is given in the text.

Table 1: A layout of a Lorenz curve with deciles

i	0	1	2	3	4	5
p_i	0	20	40	60	80	100
L_i	$L_0 = 0$	L_{20}	L_{40}	L_{60}	L_{80}	$L_{100} = 100$
$\tilde{L}_i$	$\tilde{L}_0 = 0$	$\tilde{L}_{20}$	$\tilde{L}_{40}$	$\tilde{L}_{60}$	$\tilde{L}_{80}$	$\tilde{L}_{100}$

We try to generalize Golden´s method in the following way. If the Lorenz curves are given in deciles, then Golden´s transformation should be $\tilde{L}_i = L_{i-1} + 10$ and if the p_i´s are not equidistant, then one has to define $\tilde{L}_i = L_{i-1} + p_i - p_{i-1}$. Following Golden´s rule, these processes have to continue until $\tilde{L}_i < L_i$. We then introduce $q = (p_i - \tilde{L}_i)/100$ and $G_G = 50q(3-q)$.

In many empirical situations, the income distribution $F(x)$ is given in grouped tables. If the mean of or total incomes in the groups are known, the cumulative distribution can be considered as a Lorenz curve, but the subintervals are usually not of constant length. The trapezium rule holds, but it still yields a positive bias for the area and negative bias for the Gini coefficient.

An obviously better alternative is to approximate the Lorenz curve with Lagrange´s interpolation (Berrut & Trefethen, 2004). We apply the Lagrange interpolation of second degree, which is a generalization of Simpson´s rule. However, we have to assume an even number of subintervals. Now the Lagrange polynomial is

$$L(p)=\sum_{i=0}^{n-1}\left(L_{2i}\frac{(p-p_{2i+1})(p-p_{2i+2})}{(p_{2i}-p_{2i+1})(p_{2i}-p_{2i+2})}+ \right.$$
$$\left. +L_{2i+1}\frac{(p-p_{2i+2})(p-p_{2i})}{(p_{2i+1}-p_{2i+2})(p_{2i+1}-p_{2i})}+L_{2i+2}\frac{(p-p_{2i+1})(p-p_{2i})}{(p_{2i+2}-p_{2i+1})(p_{2i+2}-p_{2i})}\right) \quad (3)$$

This approximate polynomial must be integrated in order to obtain an estimate of the area under the Lorenz curve. This attempt is a generalization of Simpson´s rule for cases with subintervals of varying lengths.

The comparison between different estimation methods is in general difficult to perform. These difficulties are mainly caused by the fact that the true Gini coefficient is unknown, but sometimes, where more detailed studies have already resulted in very accurate estimates, the comparisons are possible. Some authors (e.g., Gastwirth, 1972; Mehran, 1975; McDonald & Ransom, 1981; Rigo, 1985; Giorgi & Pallini, 1987) have introduced interval estimates, but these are often rather broad and it is still difficult to identify the best method. Such comparison problems are eliminated if the numerical estimations are applied to theoretical distributions.

Needleman ((1978) stated that as the Lorenz curve is convex, the trapezium approximation is always greater than the actual area under the curve, so that the estimate based on this approximation is always less than the actual value of the coefficient. Furthermore, he noted that most authors using the trapezium approximation indicate that they are aware of the bias involved, but either assume the error so small as to be insignificant, or else use a large number of intervals in the belief, usually justified, that the bias will then be negligible. Needleman´s own study started from the trapezium estimate of the Gini coefficient G_L. He then introduced an improved upper estimate, G_U. His final estimate follows the "two-thirds rule", that is $G=\frac{G_L}{3}+\frac{2G_U}{3}$.

McDonald and Ransom (1981) introduced lower and upper bounds of the Gini estimates. In order to estimate the bounds of the Gini coefficient estimates,

they considered the Γ density, that is, $g(y) = \frac{\beta^{\alpha} y^{\alpha-1} e^{-y\beta}}{\Gamma(\alpha)}$ with corresponding $G = \frac{\Gamma(\alpha + ½)}{\Gamma(\alpha + 1)\sqrt{\pi}}$ and $\mu = \alpha / \beta$ and applied Monte Carlo methods.

In order to perform comparisons between the estimated and theoretical Gini coefficients we analyze classes of theoretical Lorenz curves with varying Gini coefficients. In this study we compare Gini estimates for the Pareto distributions. We define the Pareto distribution as $F(x) = 1 - x^{-\alpha}$, where $x \geq 1$ and $\alpha > 1$. The frequency function is $f(x) = \alpha x^{\alpha-1}$, the mean is $\mu = \frac{\alpha}{\alpha - 1}$, the quantiles are $x_p = \left(\frac{1}{1-p}\right)^{\frac{\alpha-1}{\alpha}}$, the Lorenz curve $L(p) = 1 - (1-p)^{\frac{\alpha-1}{\alpha}}$ and the Gini coefficient $G = \frac{1}{2\alpha - 1}$. If we consider $1.5 \leq \alpha \leq 5.0$, then the Gini coefficient satisfies the inequalities $0.111 \leq G \leq 0.500$.

3 Results

Tepping data. Gastwirth (1972) presents interval estimations of the Gini coefficient. The exact Gini estimate on Current Population Surveys (CPS) income data for 1968 was computed by Tepping, his result being 0.4014. Gastwirth´s Table 2 shows Tepping´s data grouped into a 10 subgroup Lorenz curve. He compares his Gini interval estimates with Tepping´s finding. Gastwirth (1972) considers a minimum of restrictive conditions, obtaining the interval $0.3883 < G < 0.4083$. Mehran (1975) suggests an alternative estimation method, obtaining the interval estimate $0.3883 < G < 0.4087$. The grouping limits in Table 2 are not equidistant and one cannot apply Simpson´s rule. Applying the trapezium rule yields 0.3883 and the negative bias is apparent. The Lagrange rule yields 0.4033 and the modification of the Golden´s rule yields the rather inaccurate estimate 0.3740.

Lorenzen data. Lorenzen (1980) presents information about the total distribution of income for households in Germany in 1973 in his *Tabelle 2*. The Gini coefficient calculated by Lorenzen is based on data pooled in his *Tabelle 3*, which yielded 0.30. Using Lorenzen´s *Tabelle 3*, we perform a comparison of the estimates obtained based on the trapezium rule and the Lagrange rule. The available empirical data cannot yield a comparison of the accuracy of the two methods. The estimated Gini coefficient according to the trapezium rule shows negative biases compared to Lorenzen´s result, being 0.2920. The Lagrange interpolation yields the estimate 0.3486 and the modified Golden method 0.3002.

In order to analyse the accuracy of the different methods, we include some theoretical studies in this study. For the Pareto distributions presented above, we consider $1.50 \le \alpha \le 5.00$, that is, $0.1111 \le G \le 0.5000$. This G interval corresponds to the most common Gini coefficients. The results appear in Table 2 and Figure 2. Note that Simpson´s and Golden´s rules yield similar accuracy, but the trapezium rule shows the largest errors for all levels of Gini coefficients. This theoretical study indicates that Golden´s rule is not uniformly better than the trapezium rule.

Table 2: The estimation of the Gini coefficient in per cent applied to the Lorenz curve for the Pareto distributions. The estimated Gini coefficients according to the trapezium rule are inaccurate and show negative biases. Simpson´s and Golden´s rules yield similar accuracy, but Golden´s is best for large Gini values.

	Estimates			Error		
G	Trapezium	Simpson	Golden	Trapezium	Simpson	Golden
11.11	10.858	11.044	11.104	-0.253	-0.067	-0.008
12.50	12.206	12.419	12.529	-0.294	-0.081	0.029
14.29	13.935	14.185	14.370	-0.350	-0.101	0.084
16.67	16.235	16.535	16.833	-0.431	-0.132	0.166
20.00	19.442	19.816	20.291	-0.558	-0.184	0.291
25.00	24.223	24.717	25.476	-0.777	-0.283	0.476
33.33	32.102	32.820	34.026	-1.232	-0.513	0.693
50.00	47.481	48.730	50.317	-2.519	-1.270	0.317

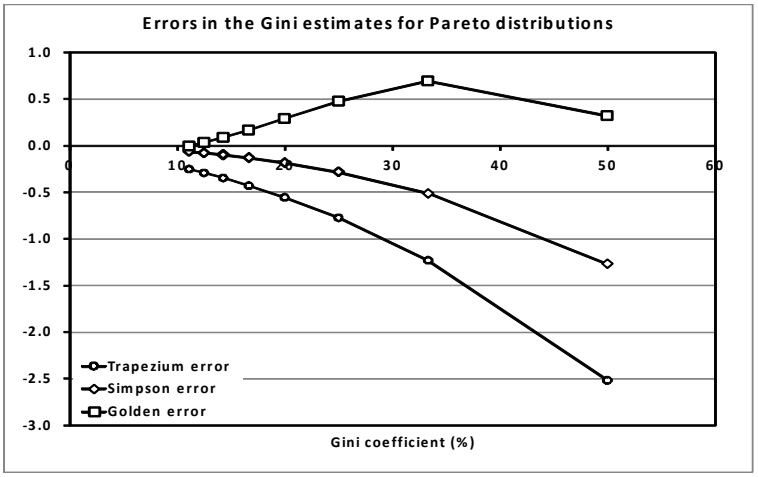

Figure 2: Estimation errors in the Gini coefficients estimated by the trapezium, Simpson´s, and Golden´s rules.

Note that Simpson´s and Golden´s rules yield similar accuracy, but the trapezium rule shows the largest errors.

5 Discussion

This study indicates that the biased trapezium rule is almost always inferior and shows negative biases. No method however is uniformly optimal. Note that Simpson´s and Golden´s rules yield similar accuracy. Golden´s method is usually of medium quality, but its accuracy fluctuates.

ACKNOWLEDGEMENTS. This work was supported in part by a grant from the Magnus Ehrnrooth Foundation.

References

[1] J.–P. Berrut and L.N. Trefethen, Barycentric Lagrange interpolation, *SIAM Review*, **46**(3), (2004), 501-517.

[2] J.L. Gastwirth, The estimation of the Lorenz curve and Gini index, *Rev. Economics and Statistics*, **54**, (1972), 306-316.

[3] G.M. Giorgi and A. Pallini, About a general method for the lower and upper distribution-free bounds on Gini´s concentration ratio from grouped data, *Statistica*, **47**, (1987), 171-184.

[4] J. Golden, A simple geometric approach to approximating the Gini coefficient. *J. Economic Education*, **39**(1), (2008), 68-77.

[5] G. Lorenzen, Was ist ein „echtes“ Konzentrationsmaß? *Allgemeines Statistisches Archiv,* **4**, (1980), 390-400.

[6] J.B. McDonald and M. R. Ransom, An analysis of the bounds for the Gini coefficient, *Journal of Econometrics,* **17**, (1981), 177-188.

[7] F. Mehran, Bounds on the Gini index based on observed points of the Lorenz curve, *J Amer Statist. Assoc. JASA*, **70**, (1975), 64-66.

[8] L. Needleman, On the approximation of the Gini coefficient of concentration, *The Manchester School*, **46**, (1978), 105-122.

[9] P. Rigo, Lower and upper distribution free bounds for Gini´s concentration ratio, *Proceedings International Statistical Institute, 45th Session, Amsterdam, Contributed Papers,* Book **2**, (1985), 629-630.