Calculating a Standard Error for the Gini Coefficient: Some Further Results

Article *in* Oxford Bulletin of Economics & Statistics · February 2004

DOI: 10.1111/j.1468-0084.2004.00086.x · Source: RePEc

David E. A. Giles
University of Victoria

OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 66, 3 (2004) 0305-9049

PRACTITIONERS' CORNER

Calculating a Standard Error for the Gini Coefficient: Some Further Results*

DAVID E. A. GILES

Department of Economics, University of Victoria, Victoria, BC, Canada V8W 2Y2 (e-mail: dgiles@uvic.ca)

Abstract

Several authors have suggested using the jackknife technique to approximate a standard error for the Gini coefficient. It has also been shown that the Gini measure can be obtained simply from an artificial ordinary least square (OLS) regression based on the data and their ranks. We show that obtaining an exact analytical expression for the standard error is actually a trivial matter. Further, by extending the regression framework to one involving seemingly unrelated regressions (SUR), several interesting hypotheses regarding the sensitivity of the Gini coefficient to changes in the data are readily tested in a formal manner.

I. Introduction

The Gini coefficient is the most widely used measure of income inequality, but it is usually reported without acknowledging the fact that it is actually a sample statistic with a sampling variance, and so a standard error should be reported. This has long been understood (e.g. Hoeffding, 1948), but Gini coefficient standard errors have rarely been reported in practice. This is because most of the formulations of this standard error that have been proposed are mathematically complex, or they are computationally intensive.[1] The latter disadvantage

*The helpful comments of the referee are gratefully acknowledged.

JEL Classification numbers: C3, C43, D31, I31.

[1]For example, see Glasser (1962), Sendler (1979), Sandstrom, Wretman and Walden (1985, 1998), and other authors cited by Ogwang (2000, p. 123).

applies, in particular, to the application of the jackknife technique to simulate a variance for the Gini coefficient, as suggested by Sandstrom, Wretman and Walden (1985, 1988) and others.[2] Recently, Karagiannis and Kovacevic (2000) and Ogwang (2000) have considered this issue. They discuss ways of reducing the computational burden associated with the jackknife approximation of the Gini coefficient's variance to a level where this method can be applied even with very large samples. Ogwang (2000) also provides a regression-based interpretation of the Gini coefficient. This note reveals that there is no need to use a jackknife approximation, and shows how the regression interpretation is helpful with regard to various hypothesis tests that are of practical interest. We illustrate our results with empirical applications.

II. Basic results

Let y be a vector of incomes, with extreme values $y_{\min}$ and $y_{\max}$, mean μ, and cumulative distribution function $F(y)$. The Gini coefficient of inequality is:

$$G = \frac{\int_{y_{\min}}^{y_{\max}} F(y)[1 - F(y)]}{\mu}. \tag{1}$$

Suppose that the observed data are in increasing order, with ith value y_i. Ogwang (2000, p. 124) notes that the Gini coefficient can be expressed as:[3]

$$G = \frac{n^2 - 1}{6n}\frac{\hat{\beta}}{\bar{y}}, \tag{2}$$

where $\bar{y}$ is the sample arithmetic mean of y, $\hat{\beta}$ is the OLS estimator of β in the model

$$y_i = \alpha + \beta i + \varepsilon_i, \tag{3}$$

and the ε_i's are zero-mean, independent, and homoskedastic errors. It is clear from equation (2) that the randomness of the Gini coefficient is determined by the sampling distributions of $\hat{\beta}$ and $\bar{y}$. These two random variables are *not* independent, so using equation (2) as the basis for measuring the variability of G does not appear to be attractive.[4] Ogwang also shows that

[2]See Efron (1982), especially Chapter 3, for details of the theoretical justification for the jackknife and other related re-sampling techniques. The jackknife was first suggested by Quenouille (1949, 1956) as a non-parametric method for estimating bias, and it was extended by Tukey (1958) to the problem of estimating varaince. Yitzhaki (1991) discusses the application of the jackknife to a range of measures related to the Gini index.

[3]See also, Lerman and Yitzhaki (1984) and Shalit (1985).

[4]In fact, by normalizing the data in equation (3) this nonlinearity can be eliminated, and so this provides a tractable basis for calculating a standard error for the Gini index. Because of the different underlying assumptions, the values of this standard error may differ from those based on equations (4) and (8) below, at least in small samples.

G can be written as:

$$G = \frac{2\hat{\theta}}{n} - 1 - \frac{1}{n}, \tag{4}$$

where $\hat{\theta}$ is the weighted least squares (WLS) estimator of θ in the model

$$i = \theta + v_i, \tag{5}$$

where the v_i's are heteroskedastic errors with variances of the form (σ^2/y_i), i.e.:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} iy_i}{\sum_{i=1}^{n} y_i}. \tag{6}$$

The randomness of G is determined linearly by the sampling distribution of $\hat{\theta}$. Ogwang's (2000) principal contribution is to use equation (4) as the basis for applying the jackknife principle to develop a standard error for G. His innovation dramatically reduces the computational burden associated with the jackknife. The key to his result is that the data are first ranked in the construction of equation (4) from equations (5) and (6).

A closer examination of Ogwang's (2000) approach reveals that using the jackknife is actually unnecessary, and the construction of an appropriate standard error for the Gini coefficient is trivial. From equation (4):

$$\mathrm{var}(G) = \frac{4\mathrm{var}(\hat{\theta})}{n^2} \tag{7}$$

and so the standard error of G is:

$$\mathrm{SE}(G) = \frac{2\mathrm{SE}(\hat{\theta})}{n}. \tag{8}$$

Of course, $\mathrm{SE}(\hat{\theta})$ comes directly from weighted least square (WLS) estimation of equation (5), or equivalently from OLS estimation of:

$$i\sqrt{y_i} = \theta\sqrt{y_i} + u_i, \tag{9}$$

where $u_i = \sqrt{y_i}v_i$. The desired standard error can be obtained directly from standard ordinary least square (OLS) regression output! This approach has been used by various authors to calculate standard errors for price indices.[5] A word of caution is needed, however. The form of heteroskedasticity assumed in equation (5) may be implausible for some data. This would be the case with many income distribution data sets, for which the relationship between

[5]For example, see Selvanathan (1991), Giles and McCann (1994), Crompton (2000). A general discussion of the stochastic approach to price index construction is provided by Clements and Izan (1987).

income and the ranks of income is convex (e.g. Pen, 1971). Accordingly, as we discuss below, it is important to test for the form of heteroskedasticity and to use standard error calculations that are robust in this respect. It should also be noted that resampling procedures are justified only in terms of their asymptotic properties.[6]

III. Numerical illustrations

We illustrate the relationship between the exact standard error given by equation (8), and its jackknife counterpart, using an artificial data set in Table 1, where we see the asymptotic convergence of the jackknife calculations and the upward bias in the jackknife Gini estimate and its standard error in finite samples.[7] In samples of 5,000 or more observations there is little difference between the standard errors from OLS/WLS and those from the jackknife. This is important in the context of income distribution studies, where several thousand observations are often used. However, the Gini coefficient is also used with other types of data involving small samples. A recent example involving industry concentration with $n = 101$ is provided by Deltas (2003, pp. 232–233). Ogwang (2000) and others have proposed that the 'exact' Gini coefficient should be used with the 'jackknife' standard error. The percentage distortion in [G/SE(G)] that would be associated with this approach is just the percentage distortion in SE(G), which is also shown in Table 1.

Next, we use consumption data for 133 countries, from the *Penn World Table*.[8] Table 2 compares the Gini coefficients and their standard errors obtained by OLS/WLS and by using the jackknife. Again, the finite-sample bias of the latter is obvious. To construct a 95% confidence interval for the Gini coefficient based on the OLS/WLS results, we can use the critical t-value of 1.978 and the standard errors.[9] For each year, this confidence interval

[6]For instance, Shao (1991) provides a detailed analysis of these properties, and establishes the weak consistency of the jackknife variance estimator. Efron and Stein (1981) prove that the jackknife variance estimator is biased upwards in small samples, so at least it provides a conservative measure.

[7]The basic data, for $n = 25$ is: $\{1\ 7\ 6\ 5\ 6\ 7\ 8\ 4\ 3\ 6\ 4\ 2\ 1\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 8\ 7\ 6\ 5\ 4\}$. The sample size is increased by assuming that the data are 'fixed in repeated samples'; i.e. if $n = 25j$, the above sample is repeated 'j' times. Accordingly, the 'exact' (OLS) Gini coefficient values shown in Table 1 are invariant to the sample size. All the calculations were undertaken with the SHAZAM (2001) econometrics package.

[8]See Summers and Heston (1995). The data were extracted using the Windows-based freeware also available at the NBER website at http://www.nber.org/pub/pwt56/. The *Penn World Tables* data set covers more countries than this, over the period 1950–92. The consumption data are in constant 1985 international prices and we have chosen a selection of recent years for which the data of interest are available for a large proportion of the countries: 1970, 1975, 1980 and 1985. The list of countries and data used in our sample are available at http://web.uvic.ca/~dgiles/ewp0202data.xls.

[9]The Student's t assumption follows if the errors in equations (5) or (9) are normally distributed. Asymptotically this will be a reasonable approximation, but the *exact* finite-sample distribution of G is another matter that we do not pursue in this paper.

TABLE 1

Gini coefficients and standard errors – artificial data

	'Exact' (OLS/WLS)		Jackknife		% Distortion
n	*G*	*SE(G)*	*G*	*SE(G)*	*in jackknife SE(G)*
25	0.2291	0.1054	0.2800	0.1125	6.7
50	0.2291	0.0738	0.2541	0.0767	3.9
100	0.2291	0.0520	0.2415	0.0533	2.5
500	0.2291	0.0231	0.2316	0.0235	1.7
1,000	0.2291	0.0164	0.2303	0.0166	1.2
5,000	0.2291	0.0075	0.2293	0.0076	0.9
10,000	0.2291	0.0052	0.2292	0.0052	0.0

TABLE 2

Gini coefficients and standard errors – PWT consumption data (133 countries)

	'Exact' (OLS/WLS)		Jackknife		Distortion in	
Year	*G*	*SE(G)*	*G*	*SE(G)*	*jackknife SE(G)*	*CV (%)*
1970	0.4705	0.0417	0.4816	0.0481	15.3	93.16
1975	0.4796	0.0405	0.4908	0.0460	13.6	93.04
1980	0.4785	0.0396	0.4897	0.0448	13.1	91.16
1985	0.4940	0.0391	0.5053	0.0441	12.8	95.20

covers the jackknife Gini estimate. Both sets of Gini coefficient estimates show an increase in consumption inequality from 1970 to 1975, a small decrease in 1980, then an increase in 1985. However, if inequality is measured by the coefficient of variation (CV), a different picture emerges. The errors in the regression model equation (5) that these results are based upon are assumed to exhibit a particular form of heteroskedasticity. When tested, this assumption cannot be rejected for any of the years considered.[10]

The OLS/WLS approach to calculating the Gini standard error also facilitates some interesting hypothesis tests that cannot be conducted readily with the jackknife approximation. For example, we can test the hypothesis that the Gini coefficient is the same in different years by stacking up single-year regressions of the form (equations (5) or (9)), using seemingly unrelated

[10]From equation (5), $\log[\mathrm{var}(v_i)] = \log(\sigma^2) - \log(y_i)$. So, in the spirit of Harvey (1976), we can fit equation (5) by OLS, then regress the logarithm of the squared residuals against an intercept and the logarithm of the consumption data, and test if the slope coefficient is -1. The corresponding *t*-statistics (and their *P*-values) for 1970, 1975, 1980 and 1985 are -0.387 (0.736), 0.017 (0.988), 0.973 (0.433) and -0.214 (0.850), respectively. Any remaining concerns about other types of heteroskedasticity can be addressed by using White's (1980) heteroskedasticity-consistent estimator of the covariance matrix, and hence of the standard errors.

TABLE 3

Gini coefficients and standard errors – PWT consumption data (SUR estimation)

Year	*G*	*SE(G)*	*% Distortion in jackknife SE(G)*	*z-tests* 1970	1975	1980
(a) Unrestricted estimation						
1970	0.3369	0.0238	102.1			
1975	0.3454	0.0231	99.1	−5.61		
1980	0.3478	0.0226	75.2	−3.94	−1.43	
1985	0.3575	0.0232	68.5	−6.74	−6.20	−6.56
(b) Restricted estimation						
1970	0.3478	0.0222	116.7			
1975	0.3552	0.0217	112.0	−5.80		
1980	0.3552	0.0217	106.5	−5.80	n.a.	
1985	0.3653	0.0213	107.0	−8.08	−7.13	−7.13

regressions (SUR) estimation, and testing the cross-equation restrictions.[11] Table 3a shows the SUR estimates of the Gini coefficient and the standard errors for our consumption data.[12] The appropriateness of SUR estimation is also clear when we test the diagonality of the model's error covariance matrix.[13] Table 3a also shows the results of testing the equality of the Gini coefficients across the equations (across years). Except for the 1975/1980 pair, we reject the hypothesis that the Gini coefficient is the same in two different years.[14] Table 3b shows the results when the 1975 and 1980 coefficients are restricted to be the same. In this case, the Gini coefficient standard errors are further reduced, and so the distortions associated with the jackknife approximation are more pronounced.

The SUR approach also allows us to consider the *significance* of the effect on the Gini measure of international consumption inequality if countries are deleted from the sample. In our sample, the USA has the highest real

[11]Each (transformed) equation of the model is of the form $i\sqrt{y_i} = \theta\sqrt{y_i} + u_i$, where the errors follow the usual SUR assumptions; i.e. they have zero mean, are serially independent, but have non-zero contemporaneous covariances (across the equations of the model).

[12]The coefficients themselves are smaller than those obtained year by year (in Table 2), and the gain in asymptotic efficiency from SUR estimation is reflected in the smaller standard errors. The latter, of course mean that the percentage distortion in the jackknife standard errors is even greater than the Table 2 results suggest.

[13]The Breusch-Pagan Lagrange multiplier test statistic is 796.12, while the corresponding likelihood ratio test statistic is 2787.50. Both statistics are asymptotically chi-square with six degrees of freedom under the null hypothesis, so we clearly reject the null of a diagonal covariance matrix.

[14]The *P*-value associated with the *z*-statistic for 1975/1980 is 15.22%. The Wald statistics for testing equivalence across *all* the years is 65.10. This statistic is asymptotically chi-square with three degrees of freedom under the null hypothesis, so the 1% critical value is 11.3449, and the *P*-value is essentially zero. We strongly reject this null hypothesis.

TABLE 4

Tests for robustness of Gini coefficient (restricted SUR estimation)

	Omit USA		*Omit Ethiopia*		*Omit USA and Ethiopia*	
Year	*G (SE(G))*	*Wald (P-value)*	*G (SE(G))*	*Wald (P-value)*	*G (SE(G))*	*Wald (P-value)*
1970	0.3505	1.9826	0.3488	4.3938	0.3514	6.3849
	(0.0223)	(0.160)	(0.0222)	(0.038)	(0.0223)	(0.041)
1975	0.3578	2.0266	0.3562	4.3382	0.3588	6.3745
	(0.0218)	(0.155)	(0.0217)	(0.037)	(0.0218)	(0.041)
1980	0.3578	2.0266	0.3562	4.3382	0.3588	6.3745
	(0.0218)	(0.155)	(0.0217)	(0.037)	(0.0218)	(0.041)
1985	0.3679	2.0566	0.3663	4.2339	0.3689	6.3001
	(0.0214)	(0.152)	(0.0213)	(0.040)	(0.0214)	(0.042)

per capita consumption in our sample, and Ethiopia has the smallest, every year. Table 4 shows the results of testing the robustness of the Gini coefficient estimates, in each year, to the deletion of one or both of these extreme sample values.[15] From Tables 3b and 4, we see that the Gini coefficient is slightly more sensitive to the omission of the USA from the sample than to the omission of Ethiopia. The Wald statistics relate to the equivalence of the Gini values before and after these omissions.[16] Dropping the USA from the sample, we *cannot* reject the null hypothesis that the Gini coefficient is unaltered, at the 15% significance level or lower. When Ethiopia is dropped, we *reject* this null hypothesis at the 5% level, although not at the 2.5% level or lower. When both countries are dropped, we *reject* the stability of the Gini coefficient at the 5% level, although not at the 4% level or lower.

IV. Concluding remarks

The Gini coefficient is the most common economic measure of inequality. A standard error is needed if confidence intervals or tests are to be constructed for this coefficient, and various authors have proposed using the jackknife technique to get a large-sample approximation for this standard error. However, because the Gini coefficient can be obtained from a simple OLS regression-based approach, the exact calculation of its standard error is actually trivial. This insight also provides the basis for constructing various tests of the robustness of the Gini coefficient to changes in the sample of data,

[15]These tests are readily implemented through the use of simple dummy variables to isolate the observations (countries) of interest.

[16]These Wald statistics are asymptotically chi-square distributed with degrees of freedom equal to the number of countries deleted from the sample.

using SUR estimation as the basis for this analysis. Such tests are not readily constructed if the jackknife methodology is used.

Final Manuscript Received: March 2004

References

Clements, K. W. and Izan, H. Y. (1987). 'The measurement of inflation: a stochastic approach', *Journal of Business and Economic Statistics*, Vol. 5, pp. 539–550.

Crompton, P. (2000). 'Extending the stochastic approach to index numbers', *Applied Economics Letters*, Vol. 7, pp. 367–371.

Deltas, G. (2003). 'The small-sample bias of the Gini coefficient: results and implications', *Review of Economics and Statistics*, Vol. 85, pp. 226–234.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Efron, B. and Stein, C. (1981). 'The jackknife estimate of variance', *Annals of Statistics*, Vol. 9, pp. 586–596.

Giles, D. E. A. and McCann, E. B. M. (1994). 'Price indices: systems estimation and tests', *Journal of Quantitative Economics*, Vol. 10, pp. 219–225.

Glasser, G. J. (1962). 'Variance formulas for the mean difference and coefficient of concentration', *Journal of the American Statistical Association*, Vol. 57, pp. 648–654.

Harvey, A. C. (1976). 'Estimating regression models with multiplicative heteroscedasticity', *Econometrica*, Vol. 44, pp. 461–465.

Hoeffding, W. (1948). 'A class of statistics with asymptotically normal distribution', *Annals of Mathematical Statistics*, Vol. 19, pp. 293–325.

Karagiannis, E. and Kovacevic, M. (2000). 'A method to calculate the jackknife variance estimator for the Gini coefficient', *Oxford Bulletin of Economics and Statistics*, Vol. 62, pp. 119–122.

Lerman, R. I. and Yitzhaki, S. (1984). 'A note on the calculation and interpretation of the Gini index', *Economics Letters*, Vol. 15, pp. 363–368.

Ogwang, T. (2000). 'A convenient method of computing the Gini index and its standard error', *Oxford Bulletin of Economics and Statistics*, Vol. 62, pp. 123–129.

Pen, J. (1971). *Income Distribution: Facts, Theories, Policies*, Praeger, New York.

Quenouille, M. (1949). 'Approximate tests of correlation in time series', *Journal of the Royal Statistical Society, Series B*, Vol. 11, pp. 18–84.

Quenouille, M. (1956). 'Notes on bias in estimation', *Biometrika*, Vol. 43, pp. 353–360.

Sandstrom, A., Wretman, J. H. and Walden, B. (1985). 'Variance estimators of the Gini coefficient: simple random sampling', *Metron*, Vol. 43, pp. 41–70.

Sandstrom, A., Wretman, J. H. and Walden, B. (1988). 'Variance estimators of the Gini coefficient: probability sampling', *Journal of Business and Economic Statistics*, Vol. 6, pp. 113–120.

Selvanathan, E. A. (1991). 'Standard errors for Laspeyres and Paasche index numbers', *Economics Letters*, Vol. 35, pp. 35–38.

Sendler, W. (1979). 'On statistical inference in concentration measurement', *Metrika*, Vol. 26, pp. 109–122.

Shao, J. (1991). 'Second-order differentiability and jackknife', *Statistica Sinica*, Vol. 1, pp. 185–202.

Shalit, H. (1985). 'Calculating the Gini index of inequality for individual data', *Oxford Bulletin of Economics and Statistics*, Vol. 47, pp. 185–189.

SHAZAM (2001). *SHAZAM Econometrics Software, User's Reference Manual*, Version 9, McGraw-Hill, New York.

Summers, R. and Heston, A. (1995). *The Penn World Tables*, *Version 5.6*, NBER, Cambridge, MA. (http://www.nber.org/pub/pwt56/).

Tukey, J. (1958). 'Bias and confidence in not quite large samples', abstract. *Annals of Mathematical Statistics*, Vol. 29, pp. 164.

White, H. (1980). 'A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity', *Econometrica*, Vol. 48, pp. 817–838.

Yitzhaki, S. (1991). 'Calculating jackknife variance estimators for parameters of the Gini method', *Journal of Business and Economic Statistics*, Vol. 9, pp. 235–239.