OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 62, 1 (2000)
0305-9049

PRACTITIONERS CORNER

A METHOD TO CALCULATE THE JACKKNIFE VARIANCE ESTIMATOR FOR THE GINI COEFFICIENT

Elias Karagiannis and Milorad Kovacevic'

The Gini coefficient is probably the most widely used measure of income inequality. However, estimates on the standard error of the coefficient are very rarely reported and the main reason for this is that it is expensive to compute. For example, Karoly (1992) reports jackknife standard errors for a number of income inequality measures except for the Gini coefficient. She claims that the cost of such computation is prohibitive, particularly for sample sizes as large as the Current Population Survey (CPS).

The purpose of this note is to present an approach for calculating the jackknife variance estimator for the Gini coefficient that is simple and efficient to compute. The estimator may be calculated with only two passes through the data regardless of the sample size.

The usual expression of the Gini coefficient is given by the following formula:

$$G = \frac{1}{2\mu N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |y_i - y_j| \tag{1}$$

where G is the Gini coefficient, μ the mean value of the distribution, N the sample size, and y_i the income of the i^{th} sample unit.

The jackknife variance estimator for the Gini coefficient is derived by the following expression (Wolter, 1985):

$$v = \frac{N-1}{N} \sum_{i=1}^{N} (G_i - G)^2 \tag{2}$$

where v is the variance estimator and G_i the value of the Gini coefficient when the i^{th} observation is taken out of the sample.

The authors wish to thank Takis Merkouris for his valuable comments.

Equation (1) clearly shows why it is prohibitively expensive to compute the standard error of the Gini coefficient. Each time an observation is taken out of the sample, we must recompute the new mean and the summation of the absolute values of the pairwise comparisons. If the sample size is fairly large, as is the case when the sample is based on labour force survey data, then it is very costly to compute the estimate and the marginal costs of computing may not justify the accuracy gained regarding the standard error of the coefficient.

This problem may be remedied by using Sen's (1973), formula for the Gini coefficient (see also Allison, 1978):

$$G = \frac{2}{\mu N^2} \sum_{i=1}^{N} r_i y_i - \frac{N+1}{N} = \frac{2}{\mu N^2} R - \frac{N+1}{N} \tag{3}$$

where r_i is the income rank of the i^{th} observation (in ascending order). During the first pass through the data the values of G and μ are calculated. However, once μ is calculated, all μ_i's may be calculated with a second pass through the data, where μ_i is the mean income of the distribution when the i^{th} observation is taken out of the sample. The relationship between μ_i and μ is given by the expression:

$$\mu_i = \frac{1}{N-1}(N\mu - y_i) \tag{4}$$

and

$$\mu = \frac{1}{N}[(N-1)\mu_i + y_i]. \tag{5}$$

It is obvious from (4) that once μ is calculated, all μ_i's may be easily calculated on the second pass through the data.

When the i^{th} observation is dropped from the sample, equation (3) becomes:

$$G_i = \frac{2}{\mu_i (N-1)^2} \sum_{j \neq i}^{N} r'_j y_j - \frac{N}{N-1} = \frac{2}{\mu_i (N-1)^2} R'_i - \frac{N}{N-1} \tag{6}$$

where r'_j is the new ranking of the data set when observation i is taken out of the sample. However, the term R'_i of (6) may be easily calculated from (3) as we shall see below. Thus, the full expression of (6) can be calculated from (3) and (4) with the second pass.

When an observation is taken out of the sample the initial rankings of the observations of the full sample and of the $(N-1)$ sample are related in the following manner: i) for all observations prior to y_i ranking does not change; ii) for all observations after y_i ranking decreases by 1, i.e. $r'_j = r_j - 1, j = i+1, \ldots, N$.

Hence, the difference between the terms R of equation (3) and R_i' of equation (6) is equal to:

$$R - R_i' = \sum_{i=1}^{N} r_i y_i - \sum_{j \neq i}^{N} r_j' y_j = r_i y_i + \sum_{j=i+1}^{N} y_j = r_i y_i + K_{i+1}. \tag{7}$$

Equation (7) can be rewritten as:

$$R_i' = R - r_i y_i - K_{i+1}. \tag{8}$$

This combined with equation (4) obtains the value of each G_i of equation (3) which can in turn be employed to derive the jackknife variance estimator of the Gini coefficient.

In summary, during the first pass through the data, the values of μ, R, and G are calculated along with the cumulatives K_i necessary for the calculations in the second pass, i.e. $K_N = y_N$, $K_{N-1} = K_N + y_{N-1}$, ..., $K_1 = K_2 + y_1$. Note that $\mu = K_1/N$. In the second pass, the values of μ_i are derived from (4), and the G_i's are calculated using a simplified form of (6), that is,

$$G_i = \frac{2}{\mu_i (N-1)^2}(R - r_i y_i - K_{i+1}) - \frac{N}{N-1}. \tag{9}$$

Most important, the value of the jackknife variance estimator, ν, is obtained in the second pass recursively from:

$$\nu_i = \frac{N-1}{N}(G_i - G)^2 + \nu_{i-1} \tag{10}$$

with $\nu_0 = 0$ and ν_N being the required variance estimator.

The method presented in this paper differs from that of Yitzhaki's (1991) in a number of ways[1]. First, Yitzhaki calculates the jackknife estimator not of the conventional Gini coefficient but the covariance presentation of the Gini mean difference (GMD). The relationship between the covariance and the Gini coefficient is as follows (see also Lerman and Yitzhaki, 1985, p. 152):

$$G = \frac{2\, cov[y,\, F(y)]}{\mu} \tag{11}$$

where $F(y)$ is the cumulative distribution of income. The formula for the covariance is:

$$cov[y,\, F(y)] = \frac{1}{N(N-1)} \sum_{i=1}^{N} r_i (y_i - \mu). \tag{12}$$

[1]We thank a referee of this journal for drawing our attention to Yitzhaki's paper.

It is the jackknife estimator of the variance of the above formula that Yitzhaki estimates and not of the conventional Gini coefficient.[2]

The method presented here computes the Gini coefficient with the first pass through the data. Finally, our formula for the jackknife estimator, (10), is more conservative (see Wolter, 1985, p. 156) since it uses the value of the Gini from the full sample.

By using Sen's formula for the Gini coefficient, it has been shown that one can derive the jackknife variance estimator of the Gini coefficient by two passes through the data. This procedure is simple to use and does not require extensive resources. Programs can easily be written using this algorithm and a SAS program is available from the authors.

Human Resources Development Canada; Statistics Canada

Date of Receipt of Final Manuscript: February 1999

References

Allison, P. D. (1978). 'Measures of Inequality', *American Sociological Review*, Vol. 43, pp. 865–80.

Karoly, L. (1992). 'Changes in the Distribution of Individual Earnings in the United States: 1967–1986', *The Review of Economics and Statistics*, Vol. LXXIV, pp. 107–15.

Lerman, Robert and Yitzhaki, Shlomo. (1985). 'Income Inequality Effects by Income Source: A new Approach and Applications to the United States', *The Review of Economic and Statistics*, Vol. 67, pp. 163–68.

Sen, A. K. (1973). *On Economic Inequality*, Oxford University Press, London.

Wolter, Kirk M. (1985). *Introduction to Variance Estimation*, Springer-Verlag, New York.

Yitzhaki, Shlomo. (1991). 'Calculating Jackknife Variance Estimators for Parameters of the Gini Method', *Journal of Business and Economic Statistics*, Vol. 9, pp. 235–39.

[2]The way Yitzhaki (1991) presents his results suggests three passes through the data. The first pass calculates μ from the full sample, the second pass calculates *cov* for the full sample and the third pass computes cov_i from *cov* and the accumulated summary statistics. The estimator of *cov* cannot be calculated from the first pass, as claimed, because it requires an estimate for μ so that the expression $\sum_{i=1}^{N} r_i(y_i - \mu)$ can be derived. However, it is possible to calculate the covariance value by the first pass if the following equivalent expression is used:

$$cov[y, F(y)] = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left(r_i - \frac{N+1}{2} \right) y_i. \tag{13}$$

With this formula, estimates of the mean and covariance can be derived from the first pass through the data. It is highly unlikely that Yitzhaki did not have the above expression in his mind but it is not clear from his paper.