CALCULATING INCOME DISTRIBUTION INDICES FROM MICRO-DATA**

STEPHEN JENKINS*

1. Introduction

SEVERAL authors have recently shown that the Gini inequality index can be rewritten as a simple linear transformation of the covariance between an income unit's rank in the income distribution and its income (Anand 1983, Lerman and Yitzhaki 1984, Nygård and Sandström 1981, Shalit 1985).[1] This is a very useful result for applied work, particularly because it coincides with a much increased availability of income micro-data sets with large numbers of observations (typically thousands). It means that, without any compromises on accuracy, researchers can calculate Gini indices easily and cheaply using widely available data management and statistical packages such as SPSS^X. They do not have to find or develop a special computer program or some approximation method (probably based on grouping of the original micro-data), as they do at present.

The message of this note is that this 'convenient covariance' approach has much wider applicability than appears to have been realised, that is, there are also analogous results for virtually all the other distributional summary statistics based on concentration curves. These include the widely-used indices of progressivity and redistributive effect proposed by Kakwani (1977), Suits (1977), and Reynolds-Smolensky (1977)-Lambert (1985), and that of horizontal inequity proposed by Atkinson (1980)-Plotnick (1981).[2]

I show this in section 2, and it will be seen that all but one of the results follow directly from a little-noted property of concentration curves. The exceptional case is the Suits index and so a new covariance result is developed for this. Section 3 discusses computational matters and section 4 provides concluding comments.

*University of Bath.

2. 'Convenient Covariance' Results

Consider a variable, y, that is some function of income, x, i.e. $y = g(x)$, and income recipients are ranked in ascending order of x. The *concentration curve* for y is defined as the share of total y received by observations with an income of x or less, $F_i[g(x)]$, graphed against the population share of those with an income no greater than x, $F(x)$. The *concentration index* is one minus twice the area under the concentration curve, i.e.

$$C_y = 1 - 2\int_0^X F_i[g(x)]dF(x), \qquad (1)$$

where $F_i[g(x)] = \int_0^x g(x)dF(x)/\tilde{g}(x)$, $\tilde{g}(x) = \int_0^X g(x)dF(x)$, and '~' denotes 'mean.' If $g(x) = x$, the concentration curve and concentration index for $g(x)$ are the Lorenz curve and Gini index, respectively, for x. Note that to calculate the Gini index for y, recipients must be ranked by y, not x; in general, $C_y \neq G_y$. (See Kakwani 1980, pp. 174–175, for the relationship between them.)

In an apparently little-known result, Kakwani (1980, p. 173) has shown that (1) can be rewritten as

$$C_y = [2/\tilde{g}(x)]\ \mathrm{cov}\ [g(x), F(x)], \qquad (2)$$

where cov[.] denotes 'covariance.' The 'convenient covariance' result for the Gini index cited at the beginning of the paper is a special case of this.

Applications

All but one of the 'convenient covariance' forms of the other distributional summary statistics cited in the Introduction follow directly from (2). Suppose that $y = g(x) = x - t(x)$, i.e. post-tax income equals 'original,' or market, income less taxes.

The *Kakwani (1977) index of tax progressivity* is defined as twice the area between the concentration curves for taxes and pre-tax income, or equivalently, the concentration index for t(x) minus the Gini index for x, i.e.

$$K = C_t - G_x, \text{ and so from (2),}$$
$$= (2/\bar{t}) \text{ cov } [t(x), F(x)]$$
$$- (2/\bar{x}) \text{ cov } [x, F(x)]. \quad (3)$$

The *Reynolds-Smolensky (1977) index of the redistributive effect of taxes,* which can also be interpreted as an index of progressivity (Lambert 1985), is defined as

$$L = G_x - G_y = [2/\bar{x}] \text{ cov } [x, F(x)]$$
$$- [2/\bar{y}] \text{ cov } [y, F(y)]. \quad (4)$$

The *Atkinson (1980)-Plotnick (1981) index* of income recipient reranking, and hence of the horizontal inequity of taxation is defined as

$$AP = (G_y - C_y)/2G_y$$
$$= \{\text{cov}[y, F^*(y)] - \text{cov}[y, F(y)]\}$$
$$\div 2\text{cov}[y, F^*(y)]. \quad (5)$$

Remember that cases are ranked in ascending order of y in calculating G_y and in order of x in calculating C_y; this distinguishes $F^*(y)$ from $F(y)$.

The *Suits (1977) index of tax progressivity* is also based on concentration curves but in this case the shares of pre-tax income, taxes and post-tax income are cumulated with respect to pre-tax income share rather than population share. I show in the Appendix that the Suits index can still be written in a 'convenient covariance' form though, viz:

$$S = (2/\bar{t}) \text{ cov } [t(x), F_i(x)]$$
$$- (2/\bar{x}) \text{ cov } [x, F(x)]. \quad (6)$$

Note the close parallel between the Kakwani and Suits indices; the key difference between them being whether the first covariance is between taxes and the cumulative share of income recipients or of pre-tax income.[3]

3. Computation

All these indices can be calculated, and the corresponding concentration curves drawn, very easily, using standard packages. These need only the capabilities listed below. To expedite implementation I list in parentheses the relevant commands in SPSS^X version 2.1 (SPSS 1986). (Analogous commands exist for, inter alia, SAS, BMDP, MINITAB, as well as many PC packages, e.g. SST.)

One needs, first, to be able to sort income recipients according to the appropriate criterion variable, pre- or post-tax income (SORT CASES).

Second, one needs to create a vector of natural numbers which corresponds to the relative ranks of each income recipient in the sorted file ($CASENUM). This rank variable, r(x), divided by the total number of recipients, n, measures F(x). For the Suits index, or for drawing concentration and Lorenz curves, one also needs to be able to calculate a cumulative income variable for recipients sorted by x (COMPUTE and LEAVE). $F_i(x)$ is this variable divided by the total income, $n\bar{x}$ (CONDESCRIPTIVE and COMPUTE). The data on $F_i(x)$ and F(x) can be used directly to draw the relevant curves (PLOT).

Third, there is the covariance calculation. This might be done directly: as the product of the corresponding correlation and the two relevant standard deviations (PEARSON CORR), or via an 'artificial' regression of y = g(x) on r(x) (REGRESSION). The point estimate of the slope coefficient from this regression is

$$b = \text{cov}[g(x), r(x)]/\text{var}[r(x)]$$
$$= 12 \text{ cov } [g(x), r(x)]$$
$$\div (n^2 - 1), \text{ and hence} \quad (7)$$
$$C_y = [(n^2 - 1)/6n](b/\bar{g}). \quad (8)$$

When calculating the Suits index, regress t(x) on $F_i(x)$. Cov[t(x), $F_i(x)$] equals the

point estimate of the slope coefficient times var[$F_i(x)$].

Concluding Remarks

Although the discussion here has been of the distributional impact of taxes, entirely analogous 'convenient covariance' expressions can be straightforwardly developed for existing indices summarizing the impact of government expenditures, and also net fiscal incidence.[4] Details are omitted here for brevity's sake.

While this note has concentrated on distributional summary statistics based on concentration curves, it is worth remembering that indices based on the Atkinson and additively decomposable Generalized Entropy families, and others, can also be calculated using standard packages (and indeed perhaps more easily than the concentration curve indices as income recipients' rankings are no longer relevant). This is because they are all based on arithmetic averages of some simple transformation of income variable(s); e.g., Atkinson's index can be written as $A = 1 - (\bar{z})^{i/(i-E)}/x$ where $z = x^{i-E}$ (COMPUTE and CONDESCRIPTIVE).[5]

The implication of all this is that computational factors should play a smaller role in 'non-programmer' analysts' choice of distributional summary statistics for large micro-data sets. Indices can be chosen more for their theoretical merits, and greater attention given to other conceptual issues such as the definition of income, income recipient etc. There is one final caveat however. The method outlined in this note completely ignores issues of sampling error (as does most of the literature to date). Taking account of these in empirical work *does* require specialized computer programs, and these are unfortunately not yet widely available.[6]

FOOTNOTES

**This paper derives from research on "Distributional change, horizontal equity, and the British tax-transfer system" financed by the ESRC (Grant number B00232123), and carried out in conjunction with Stephen Hope and Michael O'Higgins. The referees provided helpful comments.

[1]It can also be shown that the formula for the decomposition by factor components of the Gini index can be written as a function of covariances; see the Shalit and Lerman-Yitzhaki papers. And so can the 'extended Gini' (Yitzhaki 1983); see Lerman-Yitzhaki.

[2]All these indices are based on areas under or between concentration curves (though note that length-based concentration curve measures (Pfähler 1983) can also be calculated with SPSS^X). For critiques of the Kakwani and Suits indices, see inter alia Kiefer (1984), Lambert (1985), and Sykes, Smith and Formby (1987), and of the Atkinson-Plotnick index, Berliant and Strauss (1985).

[3]Similarities have been remarked upon previously, but using different formulae; see Formby, Seaks and Smith (1981).

[4]E.g. the Kienzle (1982) and Bridges (1984) indices of net fiscal progressivity are weighted sums of Suits-type indices of tax and expenditure progressivity. It should be noted that there are some important conceptual problems with these and the analogous Kakwani-type indices; see Lambert and Pfähler (1986).

[5]Indices of progressivity and redistributive effect based on the Atkinson (1970) family of inequality indices are used by Kiefer (1984). On the Generalized Entropy family of inequality indices, which includes Theil's, see Cowell and Kuga (1981). The GE family of indices of distributional change and horizontal inequity are discussed by Cowell (1980).

[6]I am aware of only one—that developed for inequality measures by Frank Cowell of ST/ICERD at the LSE.

REFERENCES

Anand, S. (1983) *Inequality and Poverty in Malaysia: Measurement and Decomposition.* Oxford: Oxford University Press.

Atkinson, A. B. (1970) "On the Measurement of Inequality," *Journal of Economic Theory* 2, 244–263.

Atkinson, A. B. (1980) "Horizontal Equity and the Distribution of the Tax Burden" in H. Aaron and M. Boskin eds., *The Economics of Taxation.* Washington: Brookings Institution.

Berliant, M. and Strauss, R. (1985) "The Horizontal and Vertical Equity Characteristics of the Federal Individual Income Tax, 1966–1977" in M. David and T. Smeeding, eds. *Horizontal Equity, Uncertainty, and Economic Well-being* NBER Studies in Income and Wealth Volume 50. Chicago: University of Chicago Press.

Bridges, B. (1984) "Post-Fisc Distributions of Income: Comment," *Public Finance Quarterly* 12, 231–240.

Cowell, F. A. (1980) "Generalized Entropy and the Measurement of Distributional Change," *European Economic Review* 13, 147–159.

Cowell, F. A. and Kuga, K. (1981) "Additivity and the Entropy Concept: An Axiomatic Approach to Inequality Measurement," *Journal of Economic Theory* 25, 131–143.

Formby, J. P., Seaks, T. G. and Smith, W. J. (1981) "A Comparison of Two New Measures of Tax Progressivity," *Economic Journal* 91, 1015–1019.

Kakwani, N. C. (1977) "Applications of Lorenz Curves in Economic Analysis," *Econometrica* 45, 719–727.
Kakwani, N. C. (1980) *Income Inequality and Poverty: Methods of Estimation and Policy Applications*. Oxford University Press.
Kiefer, D. W. (1984) "Distributional Tax Progressivity Indices," *National Tax Journal* 37, 497–513.
Kienzle, E. C. (1982) "Post-Fisc Distributions of Income: Measuring Progressivity with Application to the United States," *Public Finance Quarterly* 10, 355–368.
Lambert, P. J. (1985) "On the Redistributive Effect of Taxes and Benefits," *Scottish Journal of Political Economy* 32, 39–54.
Lambert, P. J. and Pfähler, W. (1986) "On Aggregate Measures of the Net Redistributive Impact of Taxation and Government Expenditure," Working Paper 87, Institute for Fiscal Studies, London.
Lerman, R. I. and Yitzhaki, S. (1984) "A Note on the Calculation and Interpretation of the Gini Index," *Economic Letters* 15, 363–368.
Nygård, F. and Sandström, S. (1981) *Measuring Income Inequality*. Stockholm: Almqvist and Wicksell International.
Pfähler, W. (1983) "Measuring Redistributional Effects of Tax Progressivity by Lorenz Curves," *Jahrbücher für Nationalökonomie und Statistik* 198, 237–249.
Plotnick, R. (1981) "A Measure of Horizontal Inequity," *Review of Economics and Statistics* 63, 283–288.
Reynolds, M. and Smolensky, E. (1977) *Public Expenditures, Taxes, and the Distribution of Income: the United States, 1950, 1961, 1970*. New York: Academic Press.
Shalit, H. (1985) "Calculating the Gini Index of Inequality for Individual Data," *Oxford Bulletin of Economics and Statistics* 47, 185–189.
SPSS Inc. (1986) *SPSS^X User's Guide*, second edition. New York: McGraw-Hill.
Suits, D. (1977) "Measurement of Tax Progressivity," *American Economic Review* 67, 747–752.
Sykes, D. W., Smith, J. and Formby, J. P. (1987) "On the Measurement of Tax Progressivity: An Implication of the Atkinson Theorem," *Southern Economic Journal* 53, 768–776.
Yitzhaki, S. (1983) "On an Extension of the Gini Inequality Index," *International Economic Review* 24, 617–628.

Appendix: The Suits Index in 'Convenient Covariance' Form

The Suits (1977) index is defined as

$$S = 2\int_0^x [F_i(x) - F_i[t(x)]]dF_i(x)$$

$$= 1 - 2\int_0^x F_i[t(x)]dF_i(x),$$

and integrating by parts,

$$= 2\left[\int_0^x F_i(x)dF_i[t(x)] - {}^1/_2\right]$$

and since $dF_i[t(x)] = t(x)dF(x)/\bar{t}$

$$= \frac{2}{\bar{t}}\left[\int_0^x F_i(x)t(x)dF(x) - \bar{t}\int_0^x F(x)dF(x)\right]$$

$$= \frac{2}{\bar{t}}\int_0^x [F_i(x)[t(x) - \bar{t}] + \bar{t}[F_i(x) - F(x)]]dF(x)$$

$$= \frac{2}{\bar{t}}\left[\int_0^x F_i(x)[t(x) - \bar{t}]dF(x)\right] - \left[1 - 2\int_0^x [F_i(x)dF(x)\right]$$

$$= (2/\bar{t})\ \text{cov}\ [t(x), F_i(x)] - (2/\bar{x})\ \text{cov}\ [x, F(x)],\ \text{as given in (6).}$$