The Economic Journal, **112** (*January*), 93–106. © Royal Economic Society 2002. Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

RETHINKING INEQUALITY DECOMPOSITION, WITH EVIDENCE FROM RURAL CHINA*

Jonathan Morduch and Terry Sicular

We examine inequality decompositions by income source and describe a general, regression-based approach for decomposing inequality. The approach provides an efficient and flexible way to quantify the roles of variables like education and age in a multivariate context. We illustrate the method using survey data from China. The empirical results demonstrate how sharply different conclusions can emerge for different decomposition rules. We explain how these differences reflect the treatment of equally-distributed sources of income, and we discuss implications for how results from inequality decomposition are interpreted.

A renewed interest in economic growth, political economy, and the distribution of income has again placed the economics of income inequality firmly on the academic research agenda. Relatively poor performances in Latin America have been contrasted with success stories in East Asia, and recent evidence on the changing income distributions of emerging economic giants like China has fuelled the debate (Griffin and Zhao, 1993; Knight and Li, 1997; Rozelle, 1994).

This resurgence of interest re-opens unresolved empirical questions about how to analyse income inequality and its determinants. One option is to impose as little structure as possible through non-parametric and semi-parametric methods (Deaton, 1997; Dinardo *et al.*, 1996). Researchers, however, often find it necessary to impose more structure in order to draw sharp conclusions. As a result, the most common practice is to calculate, compare, and decompose summary indices of inequality like the Gini coefficient or the variance. Inequality decomposition yields simple measures of inequality and its determinants that can be traced over time and across regions (eg, Fei *et al.*, 1978).

Decomposition by population group has been the leading approach to quantifying how education, age, etc., affect inequality. The approach begins by dividing a sample into discrete categories (eg, rural and urban residents, individuals with primary school vs. secondary or higher education) and then calculates the level of inequality within each sub-sample and between the means of the sub-samples. It can be a useful descriptive tool but has certain limitations.

First and least important, the decomposition can only be carried out over discrete categories even though some factors like age are more appropriately considered as continuous variables. Second and more important, handling

* We have benefited from discussions with Sudhir Anand, Robin Cowan, Jim Davies, Gary Fields, James Foster, Paul Gertler, Jennifer Hunt, Joseph Stern, and seminar participants at the Northeast Universities Development Consortium Conference, Harvard University, University of Western Ontario, and University of Pittsburgh, as well as from suggestions from the editor of this JOURNAL and two anonymous referees. We are grateful for the research assistance of Sarah Cook and Joseph Zveglich. Financial support was provided by the National Science Foundation under awards SES-9211260 and SES-8908438 and by the Social Science and Humanities Research Council of Canada. All views and errors are ours only.

multiple factors is often unwieldy since the number of groups increases multiplicatively with the number of categories for each factor. Indeed, as more factors and categories are added to the analysis, the number of observations in each group can diminish to the point where the within-group means and variances are highly unreliable estimates of the population moments. Third, no account can be made for the possibility that variables used to explain income inequality may themselves be partly determined by income patterns. Lack of control for endogeneity limits the decomposition to being a purely descriptive analysis.

These problems have naturally led researchers to consider ways to use regression analysis in decomposition. Use of regression estimates in inequality analysis dates at least to Oaxaca (1973) and has generated renewed interest in recent years (eg, Fields, 1998; Bourgignon *et al.*, 1998). Regression-based approaches to inequality decomposition are appealing because they overcome many of the limitations of standard decomposition by groups. For example, continuous variables are permissible, and it is possible to control for endogeneity. To date, however, work on regression-based methods of inequality has been piece-meal, with each proposed approach having different properties and using different inequality indices. For example, some add up exactly; others do not. Some use the variance of logarithms; others use the Gini or Theil indices.

The aim of the paper is to examine the underlying properties of common decomposition methods and to propose a general approach to regression-based decomposition, one that can be used with a broad class of inequality indices. The approach makes use of methods for decomposing inequality by income source (or 'factor components', Shorrocks (1982)), not by population group, but it employs estimated income flows associated with household and community characteristics rather than actual flows from labour, capital, and other direct sources of earnings. The estimated income flows are obtained using standard econometric tools (which can include corrections for endogeneity).

The approach has several important advantages over existing methods. In particular, it yields an exact allocation of contributions to the identified variables, it is general in that it can be employed with different inequality indices and decomposition rules, and it is associated with a simple procedure for deriving standard errors and confidence intervals for the estimated components of inequality.

The generality of the approach is especially valuable since, as we show, different decomposition rules have varying properties and so provide different types of information. Popular indices of aggregate inequality like the Gini coefficient, coefficient of variation, and Theil indices have well understood and very similar basic features (formally, these include satisfaction of monotonicity, scale invariance, and transfer axioms; Sen (1973)). As a result, these indices generally yield similar qualitative results. But analogues to these properties do not necessarily hold for the common forms of their decompositions by income source. The common decomposition of the Gini coefficient, for example, may indicate that wage income has a small positive influence on overall inequality, while that of the Theil-T index may show

wage income contributing substantially to inequality reduction. With these issues in mind, we begin with a reexamination of the properties of inequality measures and their decompositions by income source, highlighting the relationships between their underlying properties, patterns of empirical results, and interpretations of those results.

These issues are illustrated using data that we collected in rural Shandong Province, China, during a period when inequality widened substantially. The example from China demonstrates advantages of the regression-based decomposition method and reveals the sharp differences that can result when using alternative decomposition rules.

1. Properties of Indices and Decomposition Methods[1]

The literature contains many applications of inequality decomposition but little discussion of their interpretation or, more generally, of the underlying objective of such analysis (Shorrocks, 1999, provides a recent exception). One objective is to identify the determinants of overall inequality. Why is measured inequality 0.37, say, rather than 0.14 or 0.57? One way to answer this question is to ask how measured inequality, $I(\cdot)$, would change if income $\mathbf{y^k}$ from source k was removed from total income $\mathbf{y}$, ie, what is $I(\mathbf{y}) - I(\mathbf{y} - \mathbf{y^k})$? Examples include Danziger (1980) and Reynolds and Smolensky (1977). An alternative approach is to calculate the change in inequality that occurs when income from the k^{th} source is replaced by its mean, or $I(\mathbf{y}) - I(\mathbf{y} - \mathbf{y^k} + \mu^k\mathbf{e})$, where $\mathbf{e}$ is a vector of ones and μ^k is mean income from source k.

While the above calculations are not ideal–the sum of the parts is not guaranteed to equal the whole–they illustrate an important difference in how one approaches decomposition. The first calculation will register a reduction in inequality for sources of income that are distributed equally to all members of the population (eg, eliminating a uniform head tax). The second captures only differences from the mean, $\mathbf{y^k} - \mu^k\mathbf{e}$, and does not capture the contribution of uniform additions (or subtractions) to income because then $\mathbf{y^k} - \mu^k\mathbf{e} = \mathbf{0}$. Eliminating a head tax in this case registers as making no contribution to inequality, even though by the first calculation it will always reduce it.[2] As a result, while the second calculation provides some useful information about the sources of dispersion in income, it provides limited guidance in understanding sources of change in overall inequality over time or across regions, a frequent aim of inequality analyses.

[1] Morduch and Sicular (1998) provide an expanded discussion of theoretical issues and examples.

[2] Alternative calculations include $I(\mathbf{y^k})$, which gives the degree of inequality which would exist if all income were derived just from, say, wage income. A second, related alternative is $I[\mathbf{y^k} + (\mu - \mu^k)\mathbf{e}]$, where $\mathbf{e}$ is a vector of ones (Shorrocks, 1982). This gives the amount of inequality that would emerge if all variation were suppressed except for that of the k^{th} variable up exactly. Bourgignon *et al.* (1998) generalise the latter approach to gauge the role of changing economic structures between two points in time. For each concern k, they calculate $I[\hat{\mathbf{y}}_1^k + (\mathbf{y}_0 - \hat{\mathbf{y}}_0^k)\mathbf{e}]$ where $\hat{\mathbf{y}}_1^k$ is a prediction of the income associated with k using estimated behavioral relationships and institutions prevailing at time 1 but endowments prevailing in period 0. The estimated income component $\hat{\mathbf{y}}_0^k$ pertains only to period 0 relationships and endowments and $\mathbf{y}_0$ is actual period 0 income.

The key property can be stated formally:

DEFINITION 1. *An inequality index* $I(\mathbf{y})$ *satisfies the property of uniform additions if* $I(\mathbf{y} + \alpha\mathbf{e}) < I(\mathbf{y})$, *where* $\mathbf{y}$ *is a vector of incomes,* $\mathbf{e}$ *is a vector of ones, and* α *is a constant greater than zero.*

The property of uniform additions holds that measured inequality should fall if everyone in the population receives a positive transfer of equal size (or, conversely, that inequality should increase if everyone receives an equal, negative transfer).[3] The property of uniform additions has a direct corollary with respect to decompositions. Let there be K different sources of income, so that for any individual i total income is the sum of income from the K sources $\left(y_i = \sum_{k=1}^{K} y_i^k\right)$. Also, let $\mathbf{Y}^{\mathbf{K}}$ be the $N \times K$ matrix of source-specific income vectors $\mathbf{y}^{\mathbf{k}}$, and let s^k be the share of overall income inequality associated with income from source k. The property of uniform additions for inequality decompositions is then:

DEFINITION 2. *A decomposition method giving proportional shares of inequality* s^k *for income sources* $k = 1, \ldots, K$ *satisfies the property of uniform additions if, for an overall income distribution that is not strictly equal,* $s^k < 0$ *when* $\mathbf{y}^{\mathbf{k}} = \alpha\mathbf{e}$, *where* $\mathbf{e}$ *is a vector of ones and* α *is a constant greater than zero.*

An inequality decomposition satisfies this property if it registers strictly negative contributions to overall inequality for any income component that is equally-distributed and positive. Negative, equally-distributed components of income will register as being inequality-increasing.

Satisfaction of the property of uniform additions for an aggregate inequality index (Definition 1) does not necessarily imply that its associated decompositions satisfy Definition 2. Thus decompositions for different inequality indices, and sometimes even different decomposition rules for the same index, can yield markedly different empirical results.

Examination of the properties of several popular inequality indices and their associated decompositions demonstrates this point. We restrict our attention here to decomposition methods that, unlike the illustrative approaches above, have the attractive feature of adding up exactly (so that the sum

[3] The property of uniform additions is implied by the transfer axiom and the scale invariance axiom. The transfer axiom states that if a new distribution is obtained from another by taking income from a poorer individual and giving it to a richer individual, measured inequality should increase. This axiom, which captures a most basic idea of inequality, is satisfied by nearly all measures in common use, eg, the variance, squared coefficient of variation, Gini coefficient, Theil indices, and Atkinson index, but not by the variance of the logarithm of income (Foster and Ok, 1999). The scale invariance (homogeneity of degree zero) axiom states that if a new distribution is obtained by multiplying all incomes by a constant, measured inequality should be the same under both distributions (thus, for example, units of measurement should not affect measured inequality). All commonly-used indices, with the exception of the variance, satisfy this axiom. Scale invariance implies that income in a population can be increased proportionally while leaving overall inequality unchanged. Then the added income can be redistributed from richer to poorer people until everyone in the population ends up with an equal-sized absolute increase in income relative to the original position. The transfer axiom states that these progressive transfers reduce inequality, and consequently the property of uniform additions holds when these two axioms hold.

of the decomposition shares equals one). We also focus on the most direct and commonly used decomposition rules for each index, usually termed 'natural decomposition rules'. These rules pertain to inequality indices that can be written as a weighted sum of incomes (Shorrocks, 1982):

$$I(\mathbf{y}) = \sum a_i(\mathbf{y}) y_i. \tag{1}$$

The Gini, variance, and Theil T indices all satisfy this property. In this case, the proportional contribution of source k to overall inequality is simply

$$s^k = \frac{\sum_{i=1}^{n} a_i(\mathbf{y}) y_i^k}{I(\mathbf{y})}. \tag{2}$$

Such decompositions add up exactly, as the sum of the K proportional contributions equals one by construction.

Consider first the Gini coefficient. Where incomes are ordered so that $y_1 \leq y_2 \leq \cdots \leq y_n$, the Gini coefficient can be written as

$$I_{Gini}(\mathbf{y}) = \frac{2}{n^2 \mu} \sum_{i=1}^{n} \left(i = \frac{n+1}{2} \right) y_i. \tag{3}$$

The natural decomposition of the Gini so written gives the proportional share of inequality for source k as

$$S_{Gini}^k = \frac{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) y_i^k}{\sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) y_i}. \tag{4}$$

This turns out to be equivalent to the decomposition proposed by Fei *et al.* (1978),

$$S_{Gini}^k = \frac{\left(\frac{\mu_k}{\mu} \right) \left(\frac{\mathrm{corr}(y_i^k, i)}{\mathrm{corr}(y_i^k, i^k)} \right) I_{Gini}(\mathbf{y}^k)}{I_{Gini}(\mathbf{y})}, \tag{5}$$

where corr(.,.) refers to correlation coefficients between source k income, income ranks i for total income, and ranks i^k for source k income.

While the Gini coefficient itself satisfies the property of uniform additions (Definition 1), violation of the property with respect to the decomposition components (Definition 2) can be verified easily: if source k income is constant for all i, the proportional contribution for the source must be zero since $\mathrm{corr}(y_i^k,\ i) = \mathrm{corr}(\mu^k,\ i)$.

Similarly, the natural decomposition rule for the squared coefficient of variation (CV) and variance is

$$I_{CV}(\mathbf{y}) = I_{Var}(\mathbf{y})/\mu^2 = \frac{1}{n\mu^2} \sum_{i=1}^{n} (y_i - \mu) y_i = \frac{\mathrm{var}(\mathbf{y})}{\mu^2}, \tag{6}$$

with proportional contributions

$$S^k_{CV} = S^k_{Var} = \frac{\sum_i (y_i - \mu) y^k_i}{\sum_i (y_i - \mu) y_i} = \frac{\mathrm{cov}(\mathbf{y}^k, \mathbf{y})}{\mathrm{var}(\mathbf{y})}. \tag{7}$$

The proportional contributions for the variance and the squared CV turn out to be identical since the μ^2 terms cancel out for the squared CV. Since the covariance of income and a constant will be zero, the proportional contribution of a constant will be zero, and so, just as for the Gini, the decompositions of the variance and squared CV violate the property of uniform additions.

For an alternative formulation of the squared CV, though, the natural decomposition does satisfy the property of uniform additions.[4] Multiplying and dividing by y_i, we get an equivalent form for $I_{CV}(\mathbf{y})$:

$$I_{CV}(\mathbf{y}) = \frac{1}{n\mu^2} \sum_{i=1}^{n} \frac{(y_i^2 - \mu^2)}{y_i} y_i. \tag{8}$$

Proportional contributions are then

$$S^k_{CV} = \frac{1}{nI_{CV}(\mathbf{y})} \sum_{i=1}^{n} \frac{(y^2 - \mu^2)}{(y_i \mu^2)} y^k_i. \tag{9}$$

These shares satisfy the property of uniform additions: when $y_i > 0$ and $y^k_i = \mu^k > 0$ for all i, they simplify to $\mu^k[(1/\mu) - 1/n\sum(1/y_i)] < 0$.

The different properties of these two 'natural' decompositions of the CV reinforce the general point that the properties of decomposition components do not flow 'naturally' from the properties of the aggregate index. The illustration below using Chinese data shows that these alternative (but equally plausible) decomposition rules for the same index can yield highly dissimilar empirical results.

Finally, the property of uniform additions is also satisfied for decomposition of the Theil-T index. The Theil-T index and its natural decomposition are

$$I_{TT}(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\mu} \ln\left(\frac{y_i}{\mu}\right), \; S^k_{TT} = \frac{\frac{1}{n} \sum_{i=1}^{n} y^k_i \ln\left(\frac{y_i}{\mu}\right)}{\frac{1}{n} \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\mu}\right)}. \tag{10}$$

The role of uniform additions can be seen by considering the shares that result when $y^k_i = \mu^k$ for all i. The numerator of the share in (10) then simplifies to μ^k multiplied by the negative of the Theil-L index of overall inequality, $1/n \sum \ln(y_i/\mu)$. Given non-zero inequality, a positive, equally-distributed factor will thus always contribute a negative share in the decomposition.

All the above natural decompositions have the merits of adding up exactly and of being linked logically to the original indices. As demonstrated, however, they have different underlying properties. Thus they will yield

[4] This specification was suggested by an anonymous referee.

qualitatively different answers to the question of what determines the overall level of inequality.

2. Regression-Based Approach to Inequality Decomposition

2.1. *Integrating Inequality Decomposition with Regression Analysis*

We integrate the decomposition methods discussed above with regression analysis by applying decomposition by income source to results from estimated income equations. Our approach extends decompositions following Oaxaca (1973) that integrate regression analysis and analysis of inequality.[5] Oaxaca's concern is with sources of wage differences, while here we are interested in factors such as age, education, etc., that underlie the entire income distribution. Our approach also brings together inequality decomposition by income source and decomposition by population subgroup.

We begin with the income equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{11}$$

where $\mathbf{X}$ is an $n \times M$ matrix of independent variables with the first column given by the n-vector $\mathbf{e} = (1, 1, \ldots, 1)$, $\boldsymbol{\beta}$ is an M-vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an n-vector of residuals. The M coefficients can be estimated using appropriate econometric techniques with specification corrections as required. Predictions of per capita household income $\hat{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$ are formed using information from the entire data set. An advantage of this approach is its flexibility. Weighted least squares or quantile regressions may be used. Dividing the sample is also admissible, as are corrections for endogeneity.

Since the econometric results yield estimates of the income flows attributed to household variables, they allow us to make use of decomposition by income source (or factor income). Decomposition by income source apportions inequality to the various components of income, where the sum of these components equals total income, $y_i = \sum_{k=1}^{K} y_i^k$. In our approach $\hat{y}^{\mathbf{m}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{m}}$, the estimated income flows contributed by education, by region, by age etc., as given by the regression results, constitute the various source components of income. By construction, total income is the sum of these flows (plus the regression residual):

[5] Bourgignon *et al.* (1998) develop a promising generalisation of the Oaxaca decomposition that maps results from structural econometric equations into the Gini and Theil inequality indices. Footnote 2 above relates this method to earlier decomposition approaches and discusses their limits. Fields (1998) develops a decomposition of the variance of the logarithm of income that also incorporates regression results. This measure of inequality, however, can yield problematic results due to its violation of the transfer axiom. The alternatives correspond to the two calculations above in their treatment of uniform additions and also fail the transfer axiom (Foster and Ok, 1999), and Fields inappropriately applies the axiomatic results of Shorrocks (1982) to justify the method. In the first (1994) version of the present paper, we describe a log-linear specification of the income process mapped into a logarithmic variant of the Theil-L decomposition rule. This provides an alternative to decomposing the variance of log income, but, as demonstrated by Shorrocks (1983), the Theil-L decomposition also has unattractive properties, and so we have not pursued the specification.

$$\begin{aligned} y_i &= \sum_{m=1}^{M+1} \hat{\mathbf{y}}_i^m && \text{for all } i, \\ \text{where } \hat{\mathbf{y}}_i^m &= \hat{\beta}_m x_i^m && \text{for } m = 1, \ldots, M \\ \hat{\mathbf{y}}_i^m &= \hat{\varepsilon}_i && \text{for } m = M+1. \end{aligned} \tag{12}$$

These estimated income flows can then be used directly to calculate decomposition components for all regression variables. Following from (2), the shares take the form

$$s^m = \hat{\beta}_m \left(\frac{\sum_{i=1}^{n} a_i(\mathbf{y}) x_i^m}{I(\mathbf{y})} \right) \quad \text{for } m = 1, \ldots, M. \tag{13}$$

This formula can be applied to the decomposition of any inequality index that can be written as a weighted sum of incomes, which includes most commonly used indices.

2.2. *Computing Standard Errors*

Few empirical studies based on inequality measures calculate standard errors, although doing so is often not difficult through bootstrapping (Deaton, 1997). Since the decompositions in (13) are linear in the estimated parameters, standard errors $\sigma(\cdot)$ can be obtained simply as

$$\sigma(s^m) = \sigma(\hat{\beta}_m) \left[\frac{\sum_{i=1}^{n} a_i(\mathbf{y}) x_i^m}{I(\mathbf{y})} \right]. \tag{14}$$

For the residual, under the assumption of homoscedastic errors, $\text{var}(\varepsilon) = \sigma_\varepsilon^2$ for all i, and

$$\sigma(s^\varepsilon) = \left\{ \sigma_\varepsilon^2 \sum_{i=1}^{n} \left[\frac{a_i(\mathbf{y})}{I(\mathbf{y})} \right]^2 \right\}^{1/2}. \tag{15}$$

These standard errors provide confidence intervals for the estimated contributions of different variables to the aggregate inequality index, and their interpretation is analogous to that of standard errors for regression coefficients.

3. Illustration from Rural China

We apply the above method using data from a stratified random sample survey of 259 farm households in 16 villages in Zouping County, a relatively unexceptional rural county situated south of the Yellow River in central Shandong Province. The survey followed the households over four years,

providing 1,036 observations covering the calendar years 1990 to 1993. The survey was tailored to deliver consistent definitions of income and to provide information on a wide array of economic, social, and political variables.[6]

We begin with estimation of the determinants of income. The dependent variable is household income per capita (averaged over the four years for each household to minimise transitory fluctuations). Explanatory variables include household size, the ratio of adult workers to dependents, the number of male workers relative to the total family labour force, average education of adults (and education squared), the average age of household members, cultivated land per capita, and the extent of land fragmentation into multiple plots. A set of dummy variables for village of residence captures further geographic diversity in land quality, water conditions, distance to markets, local leadership, and paths of development.[7]

One set of variables thought to be particularly important in explaining rising inequality in China is political status and connections (Morduch and Sicular, 2000), but political status tends to be highly correlated with other household characteristics. An advantage of the regression-based decomposition method is that we can quantify the direct role of the political variables while controlling for the broad array of household and village characteristics above. Political status and connections are captured by dummy variables for the presence of a Communist Party member within the household, for the presence of a past or present cadre at the village-level or higher, and for a past or present cadre at the sub-village (*xiao zu*) level (but never above that level). Political status is also captured by the class labels given to families in the late 1940s and early 1950s by the Communist regime–landless, poor peasant, middle peasant, rich peasant and landlord. These labels were based loosely on household economic situations before the Revolution, and they influenced the political and economic treatment of households and individuals after the Revolution. We aggregate the landless and poor peasant classes and treat middle peasants as the omitted category.

The first column of Table 1 gives the coefficients from a linear earnings equation estimated by OLS on the full sample of households, and the second column gives standard errors. The remaining columns in Table 1 show the relative magnitudes and distributions of the effects of the explanatory variables on income, providing a bridge between the regression results and the decompositions that follow in Table 2. Column 3 of Table 1 gives average income shares $\hat{\beta}_m(\bar{X}^m/\bar{y})$, the fraction of mean per capita income that is given

[6] The data set contains detailed data on prices, outputs, inputs, off-farm earnings, and other sources of income, allowing calculation of income in a way that is consistent with the standard economic definition. Income is the sum of net income from household production, earnings from off-farm employment, and net transfers (including net transfers from the state, collective, and private individuals). Output retained for own consumption is valued at market prices. Income is deflated and expressed in constant 1990 prices. Household size is adjusted for absences during part of the year. Thus, the denominator of per capita income is calculated as the number of people resident for at least one month multiplied by the number of months each is present divided by twelve. Detailed descriptions of the county and survey sample are available in Sicular (1998) and Walder (1998).

[7] In the inequality decompositions below we take the mean from the village dummies and add it to the constant so that, like the residual, the village component only captures dispersion around zero.

Table 1

Regression Results and the Distributions of Income Flows from Explanatory Variables Zouping County, 1990–3

	Linear Earnings Equation			Shares of income flows by quartile				Ratio of top 25% to bottom 25%
	Estimated coefficients	Standard errors	Income shares	Bottom	Second	Third	Top	
Land per capita	89.9	80.6	11.8	18.2	28	25.6	28.2	1.5
Number of plots	−184.9*	56	−48.3	19.2	27.5	27.4	26	1.4
Household size	55.1	47.3	17.5	24.4	26.1	24.2	25.3	1
Workers per household	765.1*	232.4	40.2	25.4	23.3	24.1	27.2	1.1
Males as % of workers	710.0*	352.5	28	24.2	26.1	23.5	26.2	1.1
Average age of adults	−7.1	7.6	−19.9	23.7	25.5	25.5	25.2	1.1
Average education of adults	315.9*	116.8	129.7	24.6	23.5	23.5	28.4	1.2
Education squared	−26.1*	10	−63.2	25.4	22.1	22.4	30.1	1.2
Communist Party member	38.5	120.2	0.2	14.6	20.1	17.1	48.2	3.3
Cadre, village or higher	301.8*	121.6	2.4	23.1	18.1	24.2	34.6	1.5
Sub-village cadre	103.1	171.8	0.9	11.7	33.3	12.5	42.4	3.6
Landless/poor class	−163.0*	84.8	−7	29.8	25.6	25.1	19.6	0.7
Rich peasant class	−199.9	208.1	−0.4	7.8	29.6	37.2	25.4	3.3
Landlord class	53.4	186.5	0.3	53.9	27.7	2.4	16	0.3

Notes: The OLS earnings equation is estimated using four-year averages of income for 256 households, with population weights and village-level fixed effects. The estimated constant term is 837.5 with a standard error of 629.2. The adjusted R^2 is 0.43. * Indicates statistical significance at the 95% level of confidence.

by the mean value of each variable multiplied by its estimated coefficient from the earnings equation. Notably, none of the political variables generates a large share of average income, but the demographic variables enter strongly, as does education, with a joint income share for the education measures of nearly 70%. More detailed discussion and interpretations of these results and of alternative specifications are given in Morduch and Sicular (2000).[8]

The next columns in Table 1 provide the distribution of the income shares of the explanatory variables across quartiles. For each quartile in the distribution, shares equal the sum of estimated income flows from each variable m over all households in the quartile, divided by the sum of flows from variable m for the entire sample. Thus, for each quartile g we calculate $\hat{\beta}_m \sum_{i \in g} x_i^m / \hat{\beta}_m \sum_{\forall i} x_i^m$. The income flows from the demographic and education variables are distributed relatively equitably, with nearly equal shares going to

[8] Qualitatively similar results from the log-linear and semi-log specifications (and associated inequality decompositions) are available in Morduch and Sicular (1998). The non-linear specifications are compatible with the varlog decomposition, but not with the Gini, squared CV/variance, and Theil decompositions.

Table 2

Decompositions of Inequality Indices Estimated Proportional Shares Zouping County, 1990–3

	Theil-T	Variance/CV	Alternative CV	Gini coefficient
Land per capita	−6.67	0.91	−5.9	1.08
Number of plots	33.72*	−1.33*	29.50*	−0.81*
Household size	−16.38	0.32	−14.37	0.09
Workers per household	−38.10*	−0.50*	−34.69*	3.56*
Males as % of workers	−26.86*	0.15*	−24.11*	0.70*
Average age of adults	18.62	0.01	16.76	−0.33
Average education of adults	−94.23*	15.73*	−81.54*	16.94*
Education squared	39.16*	−12.72*	33.08*	−13.70*
Communist Party member	0.07	0.1	0.05	0.42
Cadre, village or higher	0.35*	1.00*	0.45*	4.49*
Sub-village cadre	−0.81	−0.1	−0.8	0.16
Landless/poor class	9.24*	0.75*	8.08*	1.40*
Rich peasant class	0.07	−0.03	0.07	0.01
Landlord class	−0.72	−0.17	−0.62	−0.04
Villages	98.64*	37.99*	90.38*	46.03*
Constant	−6.90*	0	−6.15*	0.00*
Regression residual	90.62*	57.85*	89.63*	39.81*
Total	*100*	*100*	*100*	*100*

Notes: All calculations are weighted by population except for the Gini decomposition.
* Indicates statistical significance at the 95% level of confidence.

the bottom and top quartiles of the income distribution. For land and education the ratios of income flows going to the top 25% versus the bottom 25% are also fairly equal – for land 1.5, and for education 1.2. Surprisingly, flows from being a village-level cadre are relatively even (the inter-quartile range is only 1.5), while income flows associated with the other political variables are among the most uneven. None of the political variables, however, generates a large share of average income.

These estimated income flows are then mapped into decomposition shares in Table 2. Results are given for the Theil-T, squared CV/variance, 'alternative' CV specification, and Gini coefficient. All four decompositions give little weight to the political variables and strongly register the inequality-increasing effects of spatial inequalities (reflected by village characteristics) and the regression residual.[9] While the data are from just one county, the importance of village location is consistent with the substantial role of spatial segmentations found by Rozelle (1994) and Jalan and Ravallion (1997) for samples from multiple provinces in China, of Knight and Li (1997) for 1,000 households in seven villages in Hebei Province, and of Benjamin and Brandt (1997) for Manchuria (northeast China) under Japanese occupation in the 1930s.

[9] For ease of exposition, after estimating we construct a single variable which aggregates the 15 village dummy variables and centre it at zero. We then adjust the constant to account for the average sample-wide effects picked up by the village dummy variables. The delta method is then used to calculate standard errors for the aggregated village variable and the adjusted constant.

The results for the Theil-T neatly follow from and summarise the patterns described in Table 1. As expected, where income sources contribute positively to total income and are distributed evenly (workers per household, fraction male adults, education, household size, and the constant term), the Theil-T decomposition registers substantial inequality reductions. Where income sources contribute negatively and are distributed relatively evenly (age, land fragmentation, education squared), the Theil-T decomposition shows substantial inequality increases. The 'alternative' CV decomposition (which, like the Theil-T decomposition, satisfies the property of uniform additions) gives similar results in this application.

In a striking contrast, decompositions of the Gini coefficient and squared CV/variance indicate that income flows associated with education, which are distributed fairly equitably, contribute *positively* to inequality by roughly 3%, and it may be tempting to infer from these results that the education system slightly exacerbates inequalities. The Theil-T and alternative CV results show, however, that this is clearly not so, as they indicate that education strongly reduces inequality, with a net negative contribution of 50–55%. Similarly, the Gini and CV/variance decompositions suggest that demographic patterns have had little bearing at all on inequality, while the Theil-T and alternative CV decompositions show that in fact the relatively even distribution of demographic characteristics has helped to keep inequality low. These examples illustrate how satisfaction (or not) of the property of uniform additions by the decomposition rule can yield qualitatively different results with opposing policy implications.

4. Concluding Comments

Better understanding patterns of inequality is essential to better understanding patterns of economic growth and development. While describing limits to the common methods employed to analyse patterns of inequality, we have described a general, regression-based approach and illustrated some of its merits. The approach consists of (i) using econometric estimation to provide conditional expectations of income, and (ii) using the estimated income flows from variables in the earnings equations to decompose inequality by income source. The approach shares the advantages of existing methods while increasing flexibility and efficiency.

In setting out our approach we have described assumptions that drive results derived from commonly-used inequality decompositions by income source. While often overlooked, key differences in results can be traced to how the decompositions treat equally-distributed sources of income (Shorrocks, 1982). For example, the aggregate Gini coefficient falls if an income source is increased by a constant amount for all members of a population, but none of the components of the standard decomposition of the Gini are affected. In contrast, both the Theil-T index and the relevant components of its decomposition register such inequality-reducing income increases.

As a result, the standard decompositions of the Gini and Theil-T yield very different sorts of information. The Gini decomposition is best interpreted as quantifying the factors that explain observed inequality *conditional* on the value of the overall Gini coefficient. It is thus of limited use in describing causes of inequality. The Theil-T decomposition provides a better indicator of why the overall index takes its given value in the first place.

Information provided by the decomposition of the Theil-T index is thus potentially of greater use to researchers, but it is seldom used. Instead, the Gini coefficient and the squared CV are the indices of choice to decompose, following the popularity of the Gini and CV themselves. Yet the results from Gini and CV decompositions are often interpreted as if they yielded the same sort of information as the Theil decomposition, leading to potential misunderstandings of the economic processes which drive income distributions.

We have applied the regression-based decomposition method to analyse patterns of inequality in a county in northern China. The debate on inequality in China has been active, centring on three explanations for emerging inequality: (a) regional segmentations, (b) human capital accumulation, and (c) political variables (Griffin and Zhao, 1993). The relative contributions of these forces to inequality in our sample are highly sensitive to the decomposition rule used. The Theil-T decomposition shows that human capital and demographic variables have, for the most part, been strongly inequality-reducing. The Gini decomposition results, however, show that these variables contribute positively, albeit modestly, to inequality. In all decompositions, the contributions of political variables are relatively small, and the contributions of spatial characteristics are large.

While results from the Theil-T decomposition are most revealing in the present application, every rule provides potentially useful information. In the end the choice of decomposition rule or approach must follow from the specific questions that are being asked. An inappropriate decomposition rule can lead researchers and policy makers astray.

New York University
University of Western Ontario
Date of receipt of first submission: May 1998
Date of receipt of final typescript: June 2001

References

Benjamin, Dwayne and Brandt, Loren (1997). 'Land, factor markets, and inequality in rural China: historical evidence', *Explorations in Economic History*, vol. 3, pp. 460–94.

Bourgignon, François, Fournier, M., and Gurgand, M. (1998). 'Distribution, development, and education: Taiwan, 1979–1994', paper presented at LACEA Conference, Buenos Aires.

Danziger, Sheldon (1980). 'Do working wives increase family income inequality?' *Journal of Human Resources*, vol.15, pp. 444–51.

Deaton, Angus (1997). *The Analysis of Household Surveys*. Baltimore: Johns Hopkins.

Dinardo, J., Fortin, N. M. and Lemiux, T. (1996). 'Labour market institutions and the distribution of wages, 1973–1992: a semi-parametric approach', *Econometrica*, vol. 64, no. 5, pp. 1001–44.

Fei, John C. H., Ranis, Gustav and Kuo, Shirley W. Y. (1978). 'Growth and the family distribution of income by factor components', *Quarterly Journal of Economics*, vol. 92, no. 1, pp. 17–53.
Fields, Gary S. (1998). 'Accounting for income inequality and its change', Department of Economics, Cornell University, mimeo.
Foster, James and Ok, Efe (1999). 'Lorenz dominance and the variance of logarithms', *Econometrica*, vol. 67, no. 4, pp. 901–8.
Griffin, Keith, and Zhao, Renwei, eds., (1993). *The Distribution of Income in China*. New York: The MacMillan Press.
Jalan, Jyotsna and Ravallion, Martin (1997). 'Geographic poverty traps?' World Bank, mimeo.
Knight, John and Li, Shi (1997). 'Cumulative causation and inequality among villages in China', *Oxford Development Studies*, vol. 25, no. 2, pp. 149–72.
Lerman, Robert and Yitzhaki, Shlomo (1985). 'Income inequality effects by income source: a new approach and applications to the United States', *Review of Economics and Statistics*, vol. 67, no. 1, pp. 151–6.
Morduch, Jonathan and Sicular, Terry (1998). 'Rethinking inequality decomposition, with evidence from rural China', Harvard Institute for International Development Discussion Paper No. 636, May.
Morduch, Jonathan and Sicular, Terry (2000). 'Politics, growth, and inequality in rural China: does it pay to join the Party?' *Journal of Public Economics*, vol. 77, no. 1, pp. 331–56.
Oaxaca, Ronald (1973). 'Male-female wage differences in urban labour markets', *International Economic Review*, vol. 14, no. 3, pp. 693–709.
Reynolds, Morgan and Smolensky, Eugene (1977). *Public Expenditures, Taxes, and the Distribution of Income*. New York: Academic Press.
Rozelle, Scott (1994). 'Rural industrialization and increasing inequality: emerging patterns in China's reforming economy', *Journal of Comparative Economics,* vol. 19, pp. 362–91.
Sen, Amartya (1973). *On Economic Inequality*. Oxford: Oxford University Press.
Shorrocks, Anthony F. (1982). 'Inequality decomposition by factor components', *Econometrica*, vol. 50, no. 1, pp. 193–211.
Shorrocks, Anthony F. (1983). 'The impact of income components on the distribution of family incomes', *Quarterly Journal of Economics*, vol. 98, no. 2, pp. 311–26.
Shorrocks, Anthony F. (1999). 'Decomposition procedures for distributional analysis: a unified framework based on the Shapley value', Department of Economics, University of Essex, mimeo.
Sicular, Terry (1998). 'Establishing markets: the process of commercialization in agriculture', in Walder (1998).
Walder, Andrew G., ed. (1998). *Zouping in Transition: The Process of Reform in Rural China*. Cambridge, MA: Harvard University Press.