United Nations University

WIDER

World Institute for Development Economics Research

Discussion Paper No. 2002/101

Regression-based Inequality Decomposition

Pitfalls and a Solution Procedure

Guang Hua Wan*

October 2002

Abstract

This paper explores pitfalls in regression-based inequality decompositions. A simple procedure is developed for rectifying these pitfalls. The procedure does not impose any restrictions on the underlying regression model and it can be applied to any inequality measure(s). Once combined with conventional decomposition methods or the Shapley value approach of Shorrocks (1999), what is being proposed becomes a most general and powerful framework for regression-based inequality decomposition. Empirical examples are provided to demonstrate the use of the procedure, and to contrast our results with those based on recent developments of Fields and Yoo (2000) and Morduch and Sicular (2002).

Keywords: regression-based decomposition, inequality, income-generating function, China

JEL classification: O15, C43, C63

*UNU/WIDER, Helsinki

This study has been prepared within the UNU/WIDER project on New Directions in Development Economics, which is directed by Tony Addison.

UNU World Institute for Development Economics Research (UNU/WIDER) was established by the United Nations University as its first research and training centre and started work in Helsinki, Finland in 1985. The purpose of the Institute is to undertake applied research and policy analysis on structural changes affecting the developing and transitional economies, to provide a forum for the advocacy of policies leading to robust, equitable and environmentally sustainable growth, and to promote capacity strengthening and training in the field of economic and social policy making. Its work is carried out by staff researchers and visiting scholars in Helsinki and through networks of collaborating scholars and institutions around the world.

UNU World Institute for Development Economics Research (UNU/WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Camera-ready typescript prepared by Lorraine Telfer-Taivainen at UNU/WIDER
Printed at UNU/WIDER, Helsinki

ISSN 1609-5774
ISBN 92-9190-328-0 (printed publication)
ISBN 92-9190-329-9 (internet publication)

1 Introduction

For years, economists have attempted to develop the regression-based approach to inequality decomposition. Pioneers in this area include Oaxaca (1973) and Blinder (1973). Juhn, Murphy and Pierce (1993) extended the earlier work to permit decomposition of between-group difference in the *full distribution* rather than in the *mean* of income only, as in Oaxaca and Blinder. Bourguignon, Fournier and Gurgran (2001) relaxed the requirement of a linear income-generating function of Juhn et al. Clearly, these efforts were devoted to explain between-group (e.g., male versus female) differences in income distribution, not to quantify contributions of many individual determinants to total inequality.

The semiparametric and nonparametric techniques, respectively proposed by DiNardo, Fortin and Lemiux (1996) and Deaton (1997), sought to model and compare the whole distribution of income in terms of the density function. However, as is typical of many non- or semiparametric methods, they often result in less conclusive findings than economists or policymakers can hope for.

In two most recent papers, Fields and Yoo (2000) and Morduch and Sicular (2002) developed new frameworks for inequality decomposition based wholly and directly on conventional regression equations. A particular advantage of the regression-based decomposition is that it enables identification as well as quantification of root causes or determinants of inequality. The number of determinants can be arbitrary; even their proxies could be used. This is not possible with any of the conventional decomposition methods. Owing to its vast flexibility and accommodating characteristics, the regression-based approach is expected to attract much attention and gain popularity.

However, the current state-of-art in regression-based inequality decomposition has a number of limitations:

(a) Severe restrictions are imposed on the functional forms of regression models that can be used. Fields and Yoo (2000), hereafter referred to as FY, requires semilog linear income-generating function. On the other hand, Morduch and Sicular (2002), hereafter referred to as MS, requires a standard linear specification.

(b) Strigent restrictions are also imposed on how inequality can be measured. In FY or Fields and Yoo (2000), inequality must be measured over logarithm of income. Under this restriction, they measure inequality by the squared coefficient of variation (CV).[1] The CV measure is known to violate the crucial principle of transfer. Conversely, only additively decomposable measures of inequality can be used in MS or Morduch and Sicular (2002). A careful reading of the paper indicates that their contribution amounts

[1] It seems that they overclaimed or exagerated the generality of their framework. For example, it is unclear how one could apply the Atkinson measure or Theil's second measure in their framework. Certainly, if one uses the Gini index, the contribution of the residual term is bound to be 0, as shown later in this paper, even if inequality is measured over logarithm of income. This contradicts their equation (4), which always allocates $1–R^2$ to the residual term.

to proposing a decomposition of Theil's first measure of inequality only. Use of other measures is either not possible or problematic (e.g., CV). Particularly disappointing is that under their framework, the most popular Gini coefficient violates their property of uniform additions (p. 98).

(c) Finally and most importantly, there exist fundamental flaws or pitfalls in the current approaches, which, so far, have been either neglected or considered not solvable. These pitfalls, if untreated, will almost always crop up in the regression-based approach to inequality decomposition and lead to misleading results.

This paper is written to accomplish three objectives. First, it represents an early attempt to expose the pitfalls in detail so theoretical and applied economists are made aware of these problems. Second, relying on the most natural rule of decomposition of Shorrocks (1999), this paper proposes a simple yet powerful procedure for regression-based inequality decomposition, which is free from the pitfalls and limitations discussed above. Third, using a set of data from China, this paper demonstrates the use of the proposed procedure and the results are contrasted with those based on FY and MS.

2 Existing pitfalls and a solution procedure

Suppose an estimated regression equation is obtained as

$$Y = F(X) + e = \alpha + Y^*(X) + e \qquad (1)$$

where Y = income or its transformation such as Ln(income) and X = income determinants or their transformations. Other notations are self-evident. It is important to note that the above model specification is more general than all earlier studies. In fact, our proposition in this paper allows for any form of $F(X)$—being linear or highly nonlinear. Both original income and logarithm of income or other transformations of income can be used as the dependent variable.

While it is seldom, though possible, to encounter a constant as a source of income in empirical analysis of income distribution, the presence of a constant is almost a rule rather than an exception in a regression equation. Such a special source is factually known to lower (raise) total inequality if it is positive (negative). For example, a headcount tax (negative constant income) increases inequality while a headcount subsidy (positive constant income) decreases inequality. However, as shown below, existing studies either avoided confronting this problem or handled it incorrectly.

A more serious and definitely unavoidable problem arises from the presence of the residual term e, which is assumed away in conventional decompositions. Although the disturbance term or its estimated counterpart is a white noise by definition, which means that it does not affect the mean of the dependent variable in (1) nor does it affect the shape of the *empirical* Lorenze curve, its presence or absence does result in different income density functions thus determines income distribution or measured inequality. Therefore, it is necessary to disentangle and identify the contribution of the residual term. Again, existing studies either avoided confronting this problem or handled it incorrectly.

The above problems must be addressed. Otherwise, the powerful regression-based approach, alternative or complementary to the conventional methodologies, will give rise to misleading results. Consequently, the potential and real advantage of this approach will be undermined and further advance in this area will be hampered. In passing, we note that the contribution of the constant term is ignored in FY while MS did not take up these issues properly. At least, prior works have not dealt with these problems according to the most natural rule of decomposition of Shorrocks (1999) or equivalently the before-after approach recommended by Cancian and Reed (1998).

In equation (1), the constant is deliberately separated out from X and the remaining deterministic part of (1) is grouped as $Y^*(X)$ or simply Y^*. This is because we mainly focus on the constant and residual terms in this paper. We are not particularly concerned about income flows from specific factors, which are less problematic and whose contributions can be handled using traditional techniques or the Shapley value framework of Shorrocks (1999).

Let us start with the residual term by re-writing (1) as

$$Y = \hat{Y} + e \tag{2}$$

where $\hat{Y} = \alpha + Y^*(X)$ represents the deterministic part of, or the predicted value of Y, from a linear or nonlinear function (α could be 0). The pitfalls can best be demonstrated by using the Gini index as an example measure of inequality. Let C denote the concentration index, applying the Gini index operator G to both sides of (2) produces:

$$G(Y) = C(\hat{Y}) + 0 \tag{3}$$

Decomposition (3) implies that the residual or disturbance term is irrelevant or plays no role at all in affecting measured income inequality. All contributions are from the deterministic part of (1). This, of course, is not correct. In addition to early discussions, one should note that $G(Y) \neq G(\hat{Y})$ unless all $e = 0$. That is, presence or absence of the residual term does alter the measured inequality in any regression-based frameworks. Given (3), it is puzzling to note the large contribution of the residual term and the near zero but significant contribution of the constant term to the measured Gini index in MS (Table 2:103).

One way to treat the residual term is to discard it altogether. After all, the residuals are not explainable by the structural income-generating function. If this is the case, one could focus on $\hat{Y}$ and obtain further decomposition results. This, however, is not recommended. Apart from earlier arguments against such a practice, the residual term, to some extent, is sometimes viewed as representing factors or determinants other than those included in the regression model. One may not be able to analyse the contribution of non-included determinants. But, ignoring e is certainly unwise as it does contain useful information. At the very least, its contribution, once identified, can inform policymakers and others as to how much included factors can explain the overall inequality. A study which leaves 70 or 80 percent of inequality to the residual term, as in MS, could be deemed useless or has very limited value. This is so despite the fact that there may exist negative and positive contributions from included factors so a large contribution from the residual term is probable.

To reiterate, ignoring the residual term means throwing away useful information on non-included determinants of income or income distribution. Such practice also causes distortions in the decomposition results, as shown in the empirical part of this paper. It is reasonable to expect that any decomposition framework in a regression-based context must come up with a relatively high explainable proportion. A benchmark might be that the net percentage due to included factors Y^* and the constant α must be no less than the contribution by the residual term.

To account for the contribution of e, we follow Shorrocks (1999, equation 2.4) by removing e from (2) and obtain

$$\mathrm{I}(Y|e = 0) = \mathrm{I}(\hat{Y}) \tag{4.1}$$

where I stands for an inequality measure. The contribution of e to I(Y) is then given by COe:

$$\mathrm{CO}e = \mathrm{I}(Y) - \mathrm{I}(\hat{Y}) \tag{4.2}$$

The difference between I(Y) and I($\hat{Y}$) is subtle and important. For example, if the Gini index is used, I(Y) = $G(Y)$ = $C(\hat{Y})$ must be calculated with Y as the ranking variable, while I($\hat{Y}$) = $G(\hat{Y})$ must be calculated with $\hat{Y}$ as the ranking variable. This is the case despite the fact that the expected values of Y and $\hat{Y}$ are identical. The rankings by Y and $\hat{Y}$ would be equivalent if and only if there is a good enough fit of the income-generating function. Viewing from this perspective, decomposition (4) makes intuitive as well as theoretical sense. Note that CO$e \rightarrow 0$ as $e \rightarrow 0$ from all directions.

Having identified the contribution of the residual term, the next task is to disentangle the contribution to I($\hat{Y}$) made by the constant term. Of course, if no constant exists in the underlying income-generating function, this issue becomes irrelevant. Provided the presence of a constant, we can then write

$$\hat{Y} = \alpha + Y^* \tag{5}$$

As before, to demonstrate our points, we use the Gini coefficient as an example measure of inequality and apply it to both sides of (5) to obtain

$$G(\hat{Y}) = \alpha / \mathrm{E}(\hat{Y})\, C(\alpha) + \mathrm{E}(Y^*)/\mathrm{E}(\hat{Y})\, C(Y^*) = 0 + \mathrm{CO}_{Y^*} \tag{6}$$

If one allocates contributions according to (6), the constant term plays no role at all. This problem always occurs to the framework of FY no matter what measures of inequality are used. It also occurs to MS, if the Gini index or CV/variance is used as measures of inequality.

MS seemed to place the blame on particular inequality measures for this problem. This is unjustified because it is their analytical framework not the inequality measures that causes this problem. To elaborate, note that the addition of a positive (or negative) constant causes a reduction (increase) in the contribution of included factors, denoted by CO_{Y^*} in (6), as E($\hat{Y}$) becomes smaller (larger) than otherwise. In other words, the impact of α did not

disappear; it is being distributed over or absorbed by other terms in $\hat{Y}$. If one can 'squeeze out' the impact entangled in the other source(s), it is natural to attribute the impact to the constant term. Unfortunately, no early efforts were devoted to 'squeeze out' this impact.

Applying the most natural rule of Shorrocks again, we have

$$I(\hat{Y} | \alpha = 0) = I(Y^*) \tag{7.1}$$

Thus, the contribution of the constant term CO_α is

$$CO_\alpha = I(\hat{Y}) - I(Y^*). \tag{7.2}$$

It is not difficult to show that $CO_\alpha < 0$ if $\alpha > 0$, and vice versa. Needless to say, when $\alpha = 0$, $\hat{Y} = Y^*$ and $CO_\alpha = 0$.

In summary, $I(Y)$ can be decomposed into CO_e, CO_α and CO_{Y^*} (which represents contributions by various non-constant Xs). The percentage contributions are

$$PC_e = 100\,[I(Y) - I(\hat{Y})]/I(Y) \tag{8}$$
$$PC_\alpha = 100\,[I(\hat{Y}) - I(Y^*)]/I(Y) \tag{9}$$
$$PC_{Y^*} = 100\, I(Y^*)/I(Y) \tag{10}$$

It is straightforward to see that the decompositions given by (8)–(10) always add to 100 percent.

What is being proposed is very general as it is independent of inequality measures to be used. The procedure, essentially embedded in equations (4) and (7), is also independent of functional specifications of $F(X)$. Furthermore, any arbitrary transformation of the target variable is allowed as long as one is prepared to measure inequality over the transformed values, as in FY. In many cases, even if the dependent variable is transformed, inequality can still be measured over the original target variable by our procedure. In those cases, one first solves the estimated equation for the original variable (e.g., taking exponentials in case of logarithm transformations) and then applies the proposed procedure. This point will be demonstrated in the next section where empirical applications are carried out.

Two points deserve special mention. Further decomposition of $I(Y^*)$ into contributions of individual determinants Xs can be undertaken using conventional methods when possible. If not, the Shapley value approach of Shorrocks (1999) can always be employed. Also, despite the fact that we have referred to income as the target variable most of the time, our results apply to any social, development or economic variable(s) under consideration.

3 Empirical examples

For demonstration purposes, we estimate two income-generating functions for rural China and then apply the proposed decomposition procedure. They are the standard linear function of MS, and the semilog of FY. These two forms are chosen partly to facilitate comparisons of results based on our procedure and those of MS and FY. With the semilog form, inequality is initially measured over logarithm of income and then measured over

original income, after taking exponentials of the estimated equation. This could be quite interesting since the resulting income-generating function is no longer a linear function. As a consequence, neither MS nor FS is applicable. The inequality measures to be considered include the most popular Gini coefficient, the conventional squared coefficient of variation (CV/Var), Theil's first measure (Theil-T) and Theil's second measure (Theil-L), and Atkinson's measure where ε is set to 0. The two Theil's measures belong to the generalised entropy family.

Aggregate data at the region level are used, covering 30 regions and the years 1992–5. Income observations are constructed as in Wan (1997, 2001). Other data are taken from China Statistical Yearbooks and China Agricultural Yearbooks (NBS, various years). To account for heterogeneity across regions, the panel data model of Kmenta (1986) is used (see Wan and Cheng 2001 for more details). This is an alternative to the random and fixed effects models. In passing, it is noted that regional dummy variables could be used to facilitate within- and between-group decomposition in the context of regression-based approach, a topic beyond the scope of this paper.

Case 1: Linear income-generating function

The estimated linear equation is (variables are expressed on a per capita basis wherever appropriate)

$$\begin{aligned}\text{Income} = &-130.61 + 49.12\ \text{HH} + 0.56\ \text{K} + 53.83\ \text{ED} - 2.13\ \text{DEP} - 13.63\ \text{Land} \\ &\quad (-1.72) \quad (3.37) \quad (9.12) \quad (9.04) \quad (-3.51) \quad (-5.24) \\ &+ 14.11\ \text{TVE} + \text{year dummies} + \text{residual} \\ &\quad (25.27)\end{aligned}$$

Buse-$R^2 = 0.92$ Sample size = 120 SSE = 110.38

In the above and below equations, t-ratios are in parentheses, HH = household size, K = per capita capital input, ED = average years of schooling, DEP = dependency ratio, TVE = proportion of labour force employed in town and village enterprises. Other variables and terms are self-explanatory. Since income is measured on a per capita basis, and dependency ratio is included, no labour input can enter the equation. The estimation results are satisfactory in terms of signs and statistical properties. Buse-R^2 is a goodness-of-fit measure when the generalised least squares estimation method is used (Buse 1973). It differs from the usual R^2 thus the residual contribution to total inequality may not equal Buse-R^2, as in FY and MS, when CV/Var is applied to the estimated models in this paper.

The negative coefficient of the land variable needs some clarifications. In China, regions with more land are usually more backward and heavily involved in farming while land scarce regions (e.g., the Pearl Delta, the Yangtze Delta) are more affluent. Farming has been a loss-making business in China since early 1990s; reports are abundant on farmers deserting land and on cases where rural households (versus urban residents by way of household registration) are administratively forced to cultivate. Land in the 1990s can be viewed as a proxy for tax contributions. As a consequence, the negative coefficient estimate is consistent with normal expectations.

The decomposition results are presented in Table 1 (recall that we focus on the constant and residual terms only in this paper). For comparison purpose, we also include decomposition results by applying the MS and FY methods to our model. We do not consider the alternative CV of MS, because its theoretical property is unknown and it had never been used before MS.

Table 1: Contributions to income inequality: linear function (%)

Inequality	FY			MS			This Paper		
Index	Residual	Constant	Other	Residual	Constant	Other	Residual	Constant	Other
Gini	X	X	X	X	X	100.0	16.69	18.06	65.25
CV/Var	X	X	X	19.10	X	80.90	10.28	19.45	70.27
Theil-T	X	X	X	41.61	26.09	32.30	22.61	28.85	48.54
Theil-L	X	X	X	X	X	X	25.66	27.43	46.91
Atkinson	X	X	X	X	X	X	25.09	27.24	47.67

Source: Author's calculations.

Note: X = not applicable.

A number of findings emerging from Table 1 are worth noting. First, our procedure can always identify contributions by the residual term as well as the constant term no matter what inequality measure is employed. In contrast, the FY method cannot be applied at all. The MS framework can only produce results for 6 out of a total of 15 cells in Table 1. Second, the residual contributions across different measures are within a comprehensible range under our decomposition procedure. However, if MS is followed, the residual contribution associated with the Theil-T measure is rather large, larger than any other components. Thirdly, the different percentage contributions given by different inequality measures essentially depend on the sensitivities of the measures to income transfers and to random shocks to the target variable. Recall that CV/Var violates the principle of transfer. This is why the contribution of the constant term is relatively small when it is used as a measure of inequality. It appears that the Gini is relatively insensitive to uniform additions or subtractions as well. With respect to the residual term or random shocks, both the CV/Var and the Gini seem less sensitive than the other measures. Finally, the decomposition results are similar when the Atkinson and GE measures are used. This is not surprising as the Atkinson measures are equivalent to GE subject to monotonic transformations (Shorrocks and Slottje 2002).

Table 1 also makes it clear that, except the Theil-T, MS cannot be meaningfully applied. In particular, if their approach is applied to the most popular and a best measure of inequality—the Gini index (see Shorrocks and Slottje 2002, Fields 2001, and Dagnum 1990)—100 percent of inequality would always be assigned to the included non-constant variables. The constant, the omitted factors and the residual term (true random shocks) do not contribute to inequality at all. In fact, even if the estimated regression model has infinitely small explanatory power, these conclusions stand—a result too good to be true. On the other hand, the framework of FY cannot be used at all as they require a strictly semilog linear form of income-generating function.

Case 2: Semilog income-generating function with inequality measured over logarithm of income

Turning to the semilog model, the following estimated equation is obtained:

$$\text{Log (Income)} = 4.52 + 0.15\ \text{HH} + 0.002\ \text{K} + 0.15\ \text{ED} - 0.007\ \text{DEP} - 0.04\ \text{Land}$$
$$(25.55)\ (4.75) \quad (11.73) \quad (11.09) \quad (-5.42) \quad (-7.62)$$

$$+ 0.02\ \text{TVE} + \text{year dummies} + \text{residual}$$
$$(29.57)$$

Buse-$R^2 = 0.97$ Sample size = 120 SSE = 111.14

As with the linear model, the estimation results seem satisfactory in terms of signs and magnitudes of the parameter estimates. The Buse-R^2 is higher than that from the linear model although they are not compatible. A particular point to note is that the constant term is now positive, thus its contribution to total inequality, if appropriately measured, should be negative. Following FY, inequality is measured over the logarithm of income. Under this circumstance, the MS approach could be applicable depending on what inequality measures to be used. The decomposition results are tabulated in Table 2.

A most striking feature of Table 2 is that there are very large negative and large positive values. This is due to the fact that the estimate of the constant term is positive and more importantly it is large relative to the values of the dependent variable. The mean of the dependent variable is 5.98, but the constant is 4.52, amounting to over 75 percent of the former. Therefore, the contribution of the constant term, which is bound to be negative, must be quite substantial. After dividing by a small value of total inequality (when inequality is measured over logarithm of income, its value becomes much smaller), the resulting percentage value is expected to be quite large. In fact, when original income is used, the Gini index is 0.20, but when the logarithm of income is used, the total inequality as measured by the Gini index is only 0.03. Even if the contribution of the constant were –0.1, the percentage contribution would amount to –333 percent. As the CV and Gini index are less sensitive to uniform additions or subtractions, other measures are likely to produce even larger values, a phenomena apparent in Table 2.

Table 2: Contributions to inequality of log (income): semilog function (%)

Inequality	FY			MS			This Paper		
Index	Residual	Constant	Other	Residual	Constant	Other	Residual	Constant	Other
Gini	X	X	X	X	X	100	21.38	-245.05	323.67
CV/Var	25.75	X	74.25	25.75	X	74.25	15.84	-262.30	346.46
Theil-T	X	X	X	52.10	-75.16	122.93	29.78	-1067.52	1137.74
Theil-L	X	X	X	X	X	X	30.37	-1026.65	1096.28
Atkinson	X	X	X	X	X	X	30.36	-1018.20	1087.84

Source: Author's calculations.

Note: X = not applicable.

The results in the last three columns may look abnormal. As discussed earlier, this is partly because inequality was measured over the logarithm of income and partly due to the large and positive estimate of the intercept term. The abnormality will disappear once inequality is measured over the original income, based on the *same* estimated regression equation. See Case 3 below. The abnormal results are presented here to demonstrate the inadequacy of FY in measuring inequality over logarithm of the target variable. In practice, results in Table 2 should be discarded altogether in favour of Table 3, if one insists on fitting a semilog income generation function. Readers are reminded that results of FY and MS cannot be compared with ours as their frameworks are flawed. We present their results in Tables 1–3 to highlight cases where their methods are not applicable.

The flaws in MS and FY are even clearer according to Table 2. Given the fact that a positive constant must exert negative contributions to total inequality, contributions from the residual term and other variables must add to more than 100 percent. The approaches given by both MS and FY, when CV/Var is used, are not capable of producing this result—they always add to 100 percent no matter what. Meanwhile, the deficiency of MS is reflected in their Theil-T decomposition results—a majority of income inequality is accounted for by the unknown residual term. This reaffirms our conclusion made earlier. In sharp contrast, our framework produces a reasonable range of residual contributions under alternative inequality measures. Again, the many Xs in Table 2 indicate the non-applicability of both MS and FY in most cases.

Case 3: Semilog income-generating function with inequality measured over original income

Measuring inequality over logarithm of the target variable is not recommended for the obvious reason that it distorts the whole distribution picture. Many unanswered questions arise when the target variable is being transformed into logarithms. In fact, even the simplest linear transformation will distort distributions (e.g., addition or subtraction of a constant from observed income data). Nevertheless, it could be argued that a semilog or double-log income-generating function is better than the linear form in that predicted values of income from logarithm models are ensured to be non-negative.

To have the 'goods' of both worlds (one with logarithm income-generating function and one demanding inequality measurement over original income), one possibility is to estimate the income-generating function in semilog or double-log form and then take exponentials to transform the estimated equation. In doing so, inequality can be measured over original income while the advantage of logarithm income-generating functions are preserved. Doing just that with our estimated semilog function gives

$$\text{Income} = \text{Exp}\{\underset{(25.55)}{4.52} + \underset{(4.75)}{0.15\ \text{HH}} + \underset{(11.73)}{0.002\ \text{K}} + \underset{(11.09)}{0.15\ \text{ED}} - \underset{(-5.42)}{0.007\ \text{DEP}} - \underset{(-7.62)}{0.04\ \text{Land}}$$

$$+ \underset{(29.57)}{0.02\ \text{TVE}} + \text{year dummies} + \text{residual}\}$$

No conventional methods nor any of the existing regression-based techniques can be employed to conduct inequality decomposition based on the above income-generating function, if one wishes to study the contributions of individual determinants. This remains the case even if the constant and residual terms can be ignored or properly dealt with. The

only alternative is the Shapley value approach of Shorrocks (1999). It is beyond the scope of this paper to consider individual contributions given our focus on the pitfalls of regression-based approach to inequality decomposition. Nevertheless, the procedure proposed in this paper can still be applied to handle the residual and the intercept terms and the results are presented in Table 3.

Table 3: Contributions to income inequality: exponential of semilog function (%)

Inequality	FY			MS			This Paper		
Index	Residual	Constant	Other	Residual	Constant	Other	Residual	Constant	Other
Gini	X	X	X	X	X	X	17.15	0	82.85
CV/Var	X	X	X	X	X	X	7.22	0	92.78
Theil-T	X	X	X	X	X	X	19.62	0	80.38
Theil-L	X	X	X	X	X	X	24.65	0	75.35
Atkinson	X	X	X	X	X	X	24.09	0	75.91

Source: Author's calculations.

Note: X = not applicable.

Table 3 demonstrates most clearly the advantages of the proposed procedure as none of early approaches are applicable at all once the underlying income generating function takes on a nonlinear form.

The 0 contributions from the constant term are acceptable because, with this particular income generating function, the constant term acts as a scaler of total income for all income recipients. Thus, the presence or absence of the constant term should not matter at all and this is exactly what is given by the proposed procedure. In passing, it is noted that the Shapley value decomposition will always attribute 0 contributions to constant terms in regression models—a potential drawback of the Shapley value approach.

If one examines the last three columns across Tables 1–3, it is clear that the contributions of residual terms are broadly comparable. Moreover, our procedure attributes most inequality to known sources including constant and various income determinants. This ought to be the case given the reasonable quality of the estimated models. In sharp contrast, MS mostly produces large percentage contributions to the unexplainable residual term. The FY framework is severely limited not only in terms of functional form but also in terms of inequality measures. While unclear as to their claim of generality, the contributions identified in Table 2 refute their claim that measures of inequality are of no relevance when inequality is measured over logarithm of income. For example, apply the Gini coefficient will result in 100 percent contribution being allocated to the included income determinants and nothing to the residual and the constant terms. But using CV/Var always attributes 0 to the constant term, $1–R^2$ to the residual term and R^2 to the other determinants, where R^2 is the usual coefficient of determination.

4 Summary and concluding remarks

This paper explores pitfalls commonly associated with regression-based inequality decompositions. A solution procedure is proposed to rectify these pitfalls. This procedure is applicable irrespective of the form of the underlying regression function. It is also applicable irrespective of what inequality measures are to be used. Empirical results are presented to demonstrate the use of the proposed procedure and to illustrate deficiencies in the recent advances in regression-based decompositions. It is important to point out that the Shapley value decomposition of Shorrocks (1999) would always attribute 0 contribution to the constant term if this term were eliminated in the same way as other variables are. In other words, this latest development may also suffer from one of the pitfalls addressed in this paper. Nevertheless, the residual term could be treated as a normal variable in Shorrocks (1999), although our procedure seems more intuitively appealing.

As theoretically derived and empirically verified in this paper, the root problems relating to the constant and residual terms are not caused by the inequality index or indices used. Fundamentally, the problems lie in the construction of proposed decomposition methodologies. From this point of view, both Morduch and Sicular (2002) and Fields and Yoo (2000) are flawed and need to be rectified along the lines discussed in this paper.

References

Blinder, A.S. (1973) 'Wage Discrimination: Reduced Form and Structural Estimates', *Journal of Human Resources* 8:436-55.

Bourguignon, F., M. Fournier and M. Gurgand (2001) 'Fast Development with a Stable Income Distribution: Taiwan, 1979-94', *Review of Income and Wealth* 47(2):139-63.

Buse, A. (1973) 'Goodness of Fit in Generalised Least Squares Estimation', *American Statistician* 27:106-8.

Cancian, M. and D. Reed (1998) 'Assessing the Effects of Wives Earning on Family Income Inequality', *Review of Economics and Statistics* 80:73-79.

Dagnum, C. (1990) 'On the Relationship between Income and Inequality Measures and Social Welfare Functions', *Journal of Econometrics* 43:91-102.

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

DiNardo, J., N.M. Fortin and T. Lemiux (1996) 'Labour Market Institutions and the Distribution of Wages, 1973-1992: A Semi-parametric Approach', *Econometrica* 64(5):1001-44.

Fields, G. (2001) *Distribution and Development: A New Look at the Developing Worlds*, The MIT Press: Cambridge MA.

Fields, G.S. and G. Yoo (2000) 'Falling Labour Income Inequality in Korea's Economic Growth: Patterns and Underlying Causes', *Review of Income and Wealth* 46(2):139-59.

Juhn, C., K. Murphy and B. Pierce (1993) 'Wage Inequality and the Rise in Returns to Skill', *Journal of Political Economy* 101:410-42.

Kmenta, J. (1986) *Elements of Econometrics*, Prentice-Hall: New Jersey.

Morduch, J. and T. Sicular (2002) 'Rethinking Inequality Decomposition, with Evidence from Rural China', *The Economic Journal* 112:93-106.

NBS (National Bureau of Statistics) (various years) *China Agricultural Yearbook*, China Statistical Publishing House: Beijing.

NBS (National Bureau of Statistics) (various years) *China Statistical Yearbook*, China Statistical Publishing House: Beijing.

Oaxaca, R. (1973) 'Male-Female Wage differences in Urban Labour Markets', *International Economic Review* 14(3):693-709.

Shorrocks, A. (1999) 'Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value' (unpublished manuscript), Department of Economics, University of Essex.

Shorrocks, A. and D. Slottje (2002) 'Approximating Unanimity Orderings: An Application to Lorenz Dominance', *Journal of Economics* (Supplement 9):91-118.

Wan, G.H. (1997) 'Decomposing Changes in the Gini Index by Factor Components' (unpublished manuscript), Center for China Economic Research, Beijing University.

Wan, G.H. (2001) 'Changes in Regional Inequality in Rural China: Decomposing the Gini Index by Income Sources', *Australian Journal of Agricultural and Resource Economics* 43(3):361-81.

Wan, G.H. and E.J. Cheng (2001) 'Effects of Land Fragmentation and Returns to Scale in the Chinese Farming Sector', *Applied Economics* 33(2):183-94.