Economics Bulletin

Volume 29, Issue 2

Non-Hierarchical Bivariate Decomposition of Theil Indexes

Kam Ki Tang
The University of Queensland

Dennis Petrie
The University of Dundee

Abstract

This paper develops a method to conduct non-hierarchical bivariate decomposition of Theil indexes. The method has the merits that, first, it treats all variates symmetrically and therefore facilitates the comparison of inequalities associated with different variates and, second, it highlights the interaction between variates in the creation of inequality. The method is applied to measure gender and ethnic income inequality in Australia.

Citation: Kam Ki Tang and Dennis Petrie, (2009) "Non-Hierarchical Bivariate Decomposition of Theil Indexes", *Economics Bulletin*, Vol. 29 no.2 pp. 918-927.
Submitted: Dec 10 2008. **Published:** May 09, 2009.

1. Introduction

The Theil index is the most commonly used entropy measure in economic studies. A property of the entropy family is that its members can be decomposed into exhaustive and exclusive components. The decomposability of inequality measures has been discussed extensively in a number of studies, e.g. Bourguignon (1979) and Shorrocks (1980).

The additive decomposability of the Theil index allows the examination of how overall inequality is related to subgroup characters. For instance, we can decompose the Theil measures of population-wide income inequality into between-gender and within-gender inequalities. Likewise, we can decompose the Theil measures based on other stratifications, such as ethnicity. Since a population can be stratified by gender and ethnicity simultaneously, the question remains whether the Theil measures can be decomposed according to both variates *simultaneously*.

When the decomposition is hierarchical, the answer is yes. Hierarchical decomposition means that the Theil measures are decomposed first in one dimension and from this point into further sub divisions (Cowell 1985). For instance, in Panel A of Figure 1, the Theil index is decomposed first by ethnicity and then by gender. Using the traditional decomposition method, within-ethnicity inequality is further decomposed into within-ethnicity-between-gender inequality and within-ethnicity-within-gender inequality, and the latter is the same as within-ethnicity-within-gender inequality. In Panel B, the order of decomposition is reversed.

For hierarchical decompositions the order of decomposition matters: Panels A and B have only one common term – w_{GE}. This is not an issue if there is a natural hierarchical order between the variates, such as the province-city stratification in Akita (2003): as a city must be hierarchically under a province, and the decomposition is naturally done first by province and then by city. In many other cases, however, there is no natural hierarchical order, e.g. gender and ethnicity, occupation and education, and industry and region.

Considering the limitations of the hierarchical decomposition, this paper aims to develop a simple method to obtain a non-hierarchical bivariate decomposition of the Theil measures. The method has two merits as compared to hierarchical decomposition. Firstly, it treats all variates symmetrically and therefore facilitates the comparison of inequalities associated with different variates. Secondly, the method highlights the interaction between variates in the creation of inequality.

2. Hierarchical and Non-Hierarchical Decomposition of Theil

2.1 Hierarchical Decomposition

Consider the income inequality of a population of people with both genders and mixed ethnic backgrounds. The Theil-L index for the population is[1]

$$L = \sum_e \sum_g \sum_i \left(\frac{N_{egi}}{N} \right) \log \left(\frac{N_{egi} / N}{Y_{egi} / Y} \right) \qquad (1)$$

[1] The expressions for the Theil-T index can be obtained by swapping *Y* and *N*, Y_{egi} *and* N_{egi} *etc*. The discussion for Theil-T will be similar to that of Theil-L.

where e = ethnicity index, g = gender index, i = income division index,[2] N_{egi} = the size of group egi, $N = \sum_e \sum_g \sum_i N_{egi}$ = the size of the whole population, Y_{egi} = income of group egi, and $Y = \sum_e \sum_g \sum_i Y_{egi}$ = total income of the population.

The logarithmic function in (1) is a measure of the deviation of the income share of the group egi (i.e Y_{egi}/Y) from its population share (i.e. N_{egi}/N). If the group's income share is equal to its population share, it has its "fair share" of income and does not contribute to the inequality index. However, if the group's income share is smaller (bigger) than its population share, it contributes positively (negatively) to the index, with its contribution weighted by its population share. In other words, the Theil-L index is a weighted sum of the deviation of income share from population share for every group in a population. A point to emphasize here is that a negative contribution, just like a positive one, indicates the existence of inequality, as with a negative contribution there must exist a larger positive contribution. Given this, the total weighted sum of all contribution will never be negative.

In Panel A of Figure 1, the Theil index is decomposed into respectively within-ethnicity-gender (w_{EG}), within-ethnicity-between-gender $(w_E b_G)$, and between-ethnicity (b_E) inequalities:

$$
\begin{aligned}
L = {} & \sum_e \left(\frac{N_e}{N}\right) \log\left(\frac{N_e / N}{Y_e / Y}\right) && (b_E) \\
& + \sum_e \left(\frac{N_e}{N}\right) \left[\sum_g \left(\frac{N_{eg}}{N_e}\right) \log\left(\frac{N_{eg} / N_e}{Y_{eg} / Y_e}\right)\right] && (w_E b_G) \qquad (2) \\
& + \sum_e \sum_g \left(\frac{N_{eg}}{N}\right) \left[\sum_i \left(\frac{N_{egi}}{N_{eg}}\right) \log\left(\frac{N_{egi} / N_{eg}}{Y_{egi} / Y_{eg}}\right)\right] && (w_{EG})
\end{aligned}
$$

where $N_{eg} = \sum_i N_{egi}$, $N_e = \sum_g N_{eg}$, $Y_{eg} = \sum_i Y_{egi}$, and $Y_e = \sum_g Y_{eg}$.

b_E measures the inequality between different ethnic groups, $w_E b_G$ measures the inequality between males and females across all ethnic groups, and w_{EG} measures the inequality within each of the ethnic-gender groups.

In Panel B, the index is decomposed into within-gender-ethnicity inequality (w_{GE}), within-gender-between-ethnicity $(w_G b_E)$, and between-gender (b_G) inequalities, respectively:

[2] E.g. i = 1 for the lowest percentile of income distribution and i = 10 for the highest percentile.

$$
\begin{aligned}
L = & \sum_g \left(\frac{N_g}{N}\right)\log\left(\frac{N_g / N}{Y_g / Y}\right) && (b_G) \\
& + \sum_g \left(\frac{N_g}{N}\right)\left[\sum_e \left(\frac{N_{eg}}{N_g}\right)\log\left(\frac{N_{eg} / N_g}{Y_{eg} / Y_g}\right)\right] && (w_G b_E) \qquad (3) \\
& + \sum_g \sum_e \left(\frac{N_{eg}}{N}\right)\left[\sum_i \left(\frac{N_{egi}}{N_{eg}}\right)\log\left(\frac{N_{egi} / N_{eg}}{Y_{egi} / Y_{eg}}\right)\right]. && (w_{GE})
\end{aligned}
$$

b_G measures the inequality between males and females, $w_G b_E$ measures the inequality between ethnic groups across both gender groups, and w_{GE} is identical to w_{EG}.

2.2 Non-Hierarchical Decomposition

Since (2) and (3) must equate each other and $w_{GE} \equiv w_{EG}$, we can state

$$w_G b_E - b_E \equiv w_E b_G - b_G \equiv residue. \qquad (4)$$

We label this residue the "gender-ethnicity interaction inequality", $i_{GE} \equiv i_{EG}$. The reason for this will become clear later. Here we can write

$$b_E \equiv w_G b_E - i_{GE}, \qquad (5)$$

$$b_G \equiv w_E b_G - i_{GE}. \qquad (6)$$

Substituting (5) into (3) yields a non-hierarchical decomposition of the Theil index into four components:

$$L = w_{GE} + b_G + b_E + i_{GE} \qquad (7)$$

The decomposition is illustrated in Figure 2. Here b_G and b_E measure the parts of inequality that are associated with gender and ethnicity respectively, w_{GE} measures the part of inequality that is associated neither with gender nor with ethnicity, and, as shown next, i_{GE} measures the part of inequality that is associated with both gender and ethnicity.

2.3 Gender-Ethnicity Interaction Inequality

As in standard decomposition, total inequality is equal to the sum of within- and between-group inequalities:

$$
\begin{aligned}
L = & \sum_e \sum_g \left(\frac{N_{eg}}{N} \right) \log \left(\frac{N_{eg} / N}{Y_{eg} / Y} \right) \qquad (b_{GE}) \\
& + \sum_e \sum_g \left(\frac{N_{eg}}{N} \right) \left[\sum_i \left(\frac{N_{egi}}{N_{eg}} \right) \log \left(\frac{N_{egi} / N_{eg}}{Y_{egi} / Y_{eg}} \right) \right] \qquad (w_{GE}).
\end{aligned} \tag{8}
$$

where $b_{GE} \equiv b_{EG}$ measures the inequality between ethnic-gender groups.

Equating (7) and (8) give

$$b_{GE} \equiv w_E b_G + w_G b_E - i_{GE} . \tag{9}$$

Substituting (5) and (6) into this yield

$$i_{GE} \equiv b_{GE} - b_G - b_E . \tag{10}$$

Here we can express the gender-ethnicity interaction term as

$$
\begin{aligned}
i_{GE} &= \sum_e \sum_g \left(\frac{N_{eg}}{N} \right) \log \left(\frac{\sigma_{Neg}}{\sigma_{Yeg}} \right) \\
\sigma_{Neg} &= \frac{N_{eg} / N}{(N_e / N)(N_g / N)}, \sigma_{Yeg} = \frac{Y_{eg} / Y}{(Y_e / Y)(Y_g / Y)}
\end{aligned} \tag{11}
$$

N_j / N is equal to the probability that a person randomly selected from the population belongs to group $j = e, g \; or \; eg$. If the event that a person belongs to ethnic group *e* is independent of the event that a person belongs to gender group *g*, $\log(\sigma_{Neg})$ will be equal to zero; otherwise, it will be non-zero. Therefore, $\log(\sigma_{Neg})$ is a measure of the dependency of the two events, or more explicitly, the interaction between ethnicity *e* and gender *g* in the allocation of the population into the ethnicity-gender group *eg*. Similarly, $\log(\sigma_{Yeg})$ is a measure of the interaction between ethnicity *e* and gender *g* in the allocation of the income into the ethnicity-gender group *eg*. Hence i_{GE} is a weighted sum of the derivation of the interaction of *e* and *g* in the allocation of income into group *eg* from that of population.

It should be noticed the independence of ethnicity and gender in the allocation of income and population is a sufficient but not necessary condition for the interaction inequality to be equal to zero. This is because even if $\sigma_{Neg} \neq \sigma_{Yeg}$ for individual *e* and *g*, it is still possible that the weighted values of $\log(\sigma_{Neg} / \sigma_{Yeg})$ for various pairs of *e* and *g* cancel each other out, leaving the net effect on total inequality zero.

Lastly, a unique feature of i_{GE}, as against the conventional inequality components, is that it can be negative, due to its structural difference. When i_{GE} is negative, it represents the overlapping part of b_E and b_G; when it is positive, it represents the ‘gap’ between the two.

3. Labour Income Inequality in Australia

This section applies the proposed decomposition method to estimate gender and ethnic labour income inequality in Australia. The data are sourced from the 1998-99 Household Expenditure Survey (HES) (Australia Bureau of Statistics 2000). The magnitudes used are weekly gross wages and salaries. Due to data limitation, the country of birth is used as a proxy for ethnicity. The HES categorizes countries of birth into 10 regions. There are totally 218,187 observations in the sample.

Table 1 summarizes the percentage share of various decomposed items of the Theil-L measure for labour income inequality. For all ages combined, w_{EG} accounts for nearly 90 percent of the total inequality, distantly followed by b_G at around 10 percent. The value of b_E is less than one percent and i_{GE} is negligible. These indicate that while gender inequality is substantial, ethnic inequality is not as an important issue. Moreover, the bivariate decomposition shows that the interaction between ethnicity and gender has contributed little to income inequality.

Since labour income increases with experience (age), if a large amount of w_{EG} is due to the income gap between workers of different ages within each ethnic-gender group, it could disguise the inequality effects of gender and ethnicity. To control for the age effect, we break down the sample into five age groups. A noticeable result is that the share of b_E increases substantially for the last two age groups at about 1.8 percent and 4.78 percent respectively, indicating ethnic inequality is more prominent amongst more experienced workers. The share of i_{GE}, while remaining small in absolute terms for all age groups, has increased substantially in proportional terms, confirming the hypothesis about the masking effect of age on gender and ethnicity inequalities.

Furthermore, gender inequality measured by b_G is below one percent for the youngest age group of 15-24 but quickly rises through child bearing and family caring ages before starting to fall for those aged 55-64. Also, for the age group of 15-24 although the gender-ethnicity interaction inequality is very small, it is more than half the size of gender or ethnic inequality. This suggests that compared with gender and ethnic inequalities a large amount of inequality is due to the interaction between gender and ethnicity for the 15-24 year olds.

Since those who were born in Australia accounts for over 75 percent of the sample, we have experimented with first grouping all other nine regions together as a single group, and second excluding Australia from the sample. The results for these two cases are reported in the last two columns of Table 3. The results are largely intact, indicating the previous findings are robust to region grouping and to the migrant sub-sample. The only noticeable difference is that in the case of Australia against all other regions together, i_{GE} is negative, indicating that b_G and b_E overlap and thus the overlapping inequality cannot be attributed solely to either gender or ethnicity.

4. Concluding Remarks

In the above example, the gender-ethnicity interaction inequality is found to be small compared with the other inequality components, this suggests little interaction or overlap between gender and ethnicity inequalities. One may thus question the value of conducting such a decomposition; however, it should be noted that, in such a case we can approximate the value of the Theil index as $L \approx w_{EG} + b_E + b_G$, which is easy to interpret. Moreover, for other variates like occupation and education, ethnicity and region, the interaction or overlap is likely to be stronger and using a non-hierarchical decomposition could highlight the combination of variates that lead to additional inequality.

The method can be generalized to handle decompositions of dimensions greater than two and also can equally be applied to the Theil-T index. The number of interaction terms increases with the number of variates, but the merit of non-hierarchical decomposition as compared with hierarchical decomposition is also greater. If the number of variates is equal to *m*, the total number of hierarchical, asymmetric decompositions is equal to *m* factorial (i.e. *m*!). In comparison, using the non-hierarchical decomposition, we can focus on a single decomposition in which all variates are treated symmetrically, and the particularly important interactions can also be isolated.

References

Akita, T. (2003) “Decomposing regional income inequality in China and Indonesia using two-stage nested Theil decomposition method”, *The Annuals of Regional Science*, **37**, 55-77.

Australia Bureau of Statistics (2000) *Household Expenditure Survey: User Guide 1998-99*, Canberra.

Bourguignon, F. (1979) “Decomposable Income Inequality Measures” *Econometrica*, **47**, 901-20.

Cowell, F.A. (1985) “'Multilevel Decomposition of Theil's Index of Inequality: a note” *Review of Income and Wealth*, **31**, 201-5.

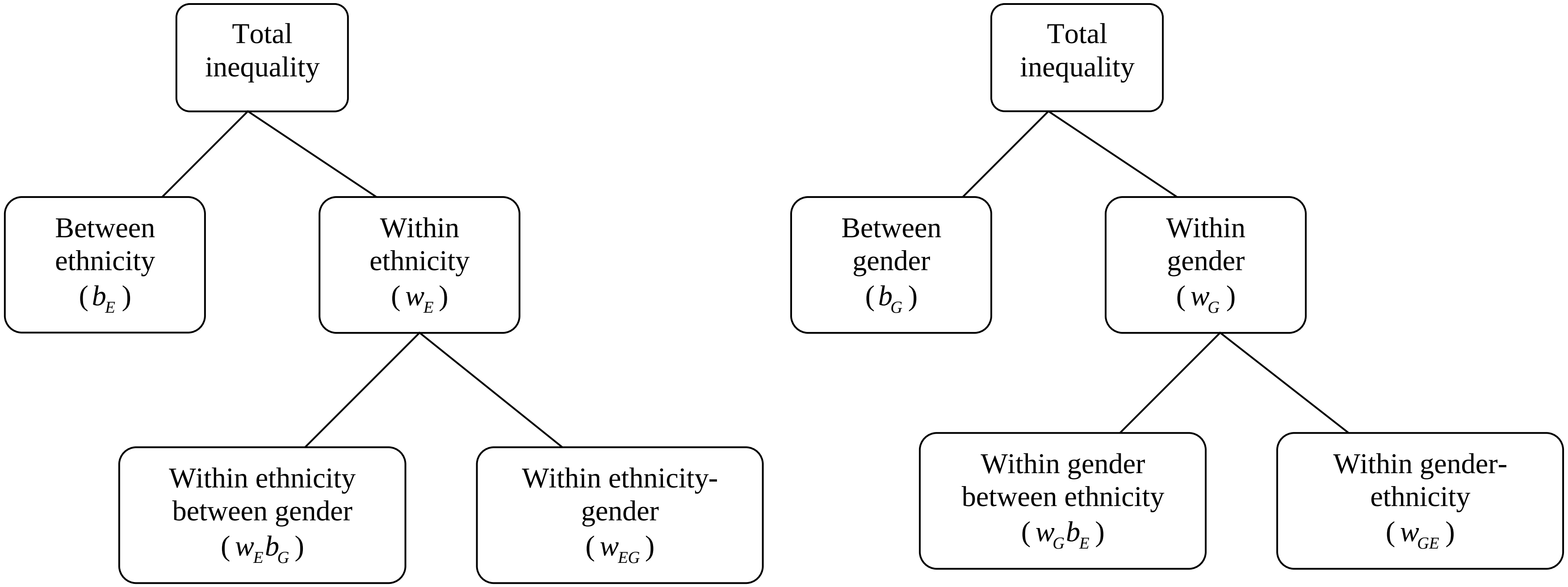
Shorrocks, A.F. (1980) “The Class of Additively Decomposable Inequality Measures” *Econometrica*, **48**, 613-25.

Table 1 Percentage Shares of Various Components of the Theil-L Index

	Ten regions						Australia vs other regions together	All regions excluding Australia
	All ages	15-24	25-34	35-44	45-54	55-64	All ages	All ages
$w_{GE} = w_{EG}$	88.81	98.02	88.77	83.42	81.90	81.64	89.29	87.75
b_E	0.88	0.54	0.85	0.77	1.80	4.78	0.52	1.58
b_G	10.30	0.84	9.89	15.32	15.72	13.31	10.30	10.10
$i_{GE} = i_{EG}$	0.01	0.60	0.49	0.50	0.59	0.27	-0.12	0.57

The ten regions are: Australia, Other Oceania and Antarctica, North-West Europe, Southern and Eastern Europe, North Africa and Middle East, South-East Asia, North-East Asia, Southern and Central Asia, Americas, and Sub-Saharan Africa.

Figure 1 Hierarchical Bivariate Decomposition of Theil Indexes



Panel A

Panel B

Figure 2 Non-hierarchical Bivariate Decomposition of Theil indexes