Decomposing Socioeconomic Inequality of Health

Guido Erreygers[a,b], Roselinde Kessels[a,c], Linkun Chen[b] and Philip Clarke[b]

[a] Department of Economics, University of Antwerp, Belgium

[b] Centre for Health Policy, University of Melbourne, Australia

[c] Flemish Research Foundation (FWO), Belgium

December 19, 2015

Abstract

We consider three different types of decomposition analysis: decomposition by health components, decomposition by subgroups, and regression-based decomposition. We show that level-dependent indices perform better than rank-dependent indices.

Keywords: Inequality measurement, Decomposition analysis, Socioeconomic inequality of health, Bivariate inequality.

JEL Classification Number: D63, I00

Adresses of the authors:

Guido Erreygers: guido.erreygers@uantwerpen.be

Roselinde Kessels: roselinde.kessels@uantwerpen.be

Linkun Chen: linda.chenlinkun@gmail.com

Philip Clarke: philip.clarke@unimelb.edu.au

Acknowledgements: This paper uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Project was initiated and is funded by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the authors and should not be attributed to either FaHCSIA or the Melbourne Institute.

1 Introduction

In the literature on the measurement of socioeconomic inequality of health, decomposition techniques have gained currency as methods to better understand, and if possible to explain, the observed levels of inequality at a given moment of time, as well as the changes in these levels over time. While some of these techniques have been borrowed or adapted from well-known results in the fields of inequality measurement (e.g., Rao, 1969; Bourguignon, 1979; Shorrocks, 1980) and labour economics (e.g., Blinder, 1973; Oaxaca, 1973), others have been developed specifically for the analysis of bivariate inequality (e.g., Wagstaff, Van Doorslaer and Watanabe, 2003; Clarke, Gerdtham and Connelly, 2003; Wagstaff, 2005b; Abul Naga and Geoffard, 2006; Makdissi, Sylla and Yazbeck, 2013). Instructive surveys of the state-of-the-art can be found in Van Doorslaer and Van Ourti (2011) and in O'Donnell, Van Doorslaer and Wagstaff (2012).

Almost all of this work is centered around rank-dependent indicators of inequality, which are the dominant measures of socioeconomic inequality of health. Recently, however, a case has been made for a shift to level-dependent indicators (Erreygers and Kessels, 2015). One of the considerations which seems particularly relevant for the choice between different indicators is how amenable they are when it comes to decomposability. The main purpose of this paper is to compare the decomposition properties of rank-dependent and level-dependent indicators of socioeconomic inequality of health. We do so by focusing on three forms of inequality decomposition:

- decomposition by health components;
- decomposition by population groups; and
- regression-based decomposition.

The first form of decomposition can be handled just as well by rank-dependent indices as by level-dependent indices. With respect to the decomposition by population subgroups, however, level-dependent indices have much nicer properties than rank-dependent indices. This may prove to be an important argument in favour of the use of level-dependent indices. With regard to regression-based decomposition, the paper suggests a new approach which aims to explain the correlation between socioeconomic conditions and health outcomes more directly than has been done so far.

2 Indices of Socioeconomic Inequality of Health

2.1 Basic Notation and Assumptions

We consider a population consisting of n individuals, where each individual i ($i = 1, 2, ..., n$) is characterized by their levels of socioeconomic achievement (y_i) and of health attainment (h_i). For convenience, we designate socioeconomic achievement as "income", but it goes without saying that socioeconomic status can be measured in a variety of ways (by consumption levels, by years of education, etc.). We assume that both y_i and h_i have well-defined lower bounds greater than or equal to 0. More specifically, each of these variables is either a ratio-scale variable with no upper bound, or a cardinal or ratio-scale variable with a well-defined upper bound. The two variables are not necessarily of the same nature. Income, for instance, is an unbounded ratio-scale variable. The health variable, by contrast, is often a bounded variable, and need not be of the ratio-scale type. The means of the variables are $\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\mu_h = \frac{1}{n}\sum_{i=1}^{n} h_i$, respectively.

We denote the rank of individual i in the income distribution as r_i. If there are no ties in the income distribution, the rank of the poorest person in society is equal to 1, the rank of the second poorest person equal to 2, etc., and the rank of the richest person equal to n. If a group of $k+1$ individuals are tied on position g, the rank of each of these individuals is equal to $g + (k/2)$. For simplicity, however, we assume there are no ties in the income distribution, and that individuals are ranked according to their incomes ($y_1 < y_2 < \ldots < y_n$). (Appendix 1 specifies how the formulas must be adapted when there are ties, and when not all individuals have the same population weight.)

2.2 Bivariate Linear Indices

The indices of bivariate inequality considered in this paper belong to a broad class of linear inequality measures. These indices are closely related to the family of linear measures of univariate inequality defined by Mehran (1976). We focus on bivariate indices where income serves as the weighting variable and health as the weighted variable. In formal terms, the absolute version of these indices takes the following form[1]:

$$I(y,h) = \frac{1}{n}\sum_{i=1}^{n} w_i(y) h_i \tag{1}$$

[1] If there is no confusion possible, we will drop the arguments and simply write I instead of $I(y,h)$, etc.

What distinguishes bivariate from univariate indices, is that the weights are determined by a different distribution from the one which is weighted. The Concentration Index, for example, defines the weights in terms of the ranks of individuals in the income distribution, and applies these weights to the distribution of health.

The relative version of our indices is obtained by dividing the index by the mean of the weighted variable:

$$I_{rel}(y,h) = \frac{I(y,h)}{\mu_h} \tag{2}$$

Other versions of the indices have been proposed in the literature, e.g. in order to deal with binary variables (Wagstaff, 2005a). For bounded variables, whether they are binary or of another type, we suggest the use of the index proposed by Erreygers (2009):

$$I_{bou}(y,h) = \frac{A}{(h_{\max} - h_{\min})} I(y,h) \tag{3}$$

where $h_{\min}$ and $h_{\max}$ stand for the lower and upper bounds of the health variable, and A is a constant. Erreygers and Van Ourti (2011) have argued that the choice for a specific version of a bivariate index should be made in accordance with the nature of the health variable under consideration.

2.3 Rank-Dependent Indices

Rank-dependent indices rely exclusively on income ranks to define the weights $w_i(y)$; no other information on the income distribution enters into the calculation of the weights. The standard Concentration Index, for instance, is characterized by a weighting function which is linear in the income ranks r_i. This weighting function may therefore be called the 'rank' function and the associated bivariate index the R index:

$$w_i^R = \frac{2r_i - n - 1}{n} \tag{4}$$

$$R = \frac{1}{n} \sum_{i=1}^{n} w_i^R h_i \tag{5}$$

The weights w_i^R steadily increase as r_i goes from 1 to n. If the group of the 'poor' is defined as those who have negative weights and the group of the 'rich' as those who have positive weights, the weighting function (4) puts the boundary between the two groups exactly in the middle of the population.

Those with ranks smaller than or equal to $n/2$ (if n is even) or smaller than or equal to $(n-1)/2$ (if n is odd) have negative weights, and those with ranks larger than or equal to $n/2+1$ (if n is even) or larger than or equal to $(n+1)/2$ (if n is odd) have positive weights. Put differently, individuals with an income below the median income have negative weights, and individuals with an income above the median positive weights. The negative weights sum to $-n/4$ and the positive weights to $n/4$.[2] This motivates the choice of $A = 4$ for the bounded version of the basic rank-dependent index, as suggested by Erreygers (2009).

2.4 Level-Dependent Indices

In contrast to rank-dependent indices, the weights of level-dependent indices are based upon income levels rather than income ranks. The basic version of the level-dependent index L proposed by Erreygers and Kessels (2015) has a weighting function which is a simple linear function of income:

$$w_i^L = \frac{y_i - \mu_y}{\mu_y} \tag{6}$$

$$L = \frac{1}{n} \sum_{i=1}^{n} w_i^L h_i \tag{7}$$

The weights w_i^L are proportional to the deviations of the incomes from the mean. If there is a high degree of income inequality, typically there will be a lot of individuals with income levels below the mean, who will therefore have negative, but in absolute terms rather small weights. On the other side of the spectrum, those who have very high incomes will have positive, and in absolute terms quite large weights. Hence, a small change in the health level of a relatively well-off individual will probably have a more pronounced influence on the index than a comparable change in the health level of an individual who is less well-off.

All the negative weights sum to $-(n/2)D$, and all the positive weights to $(n/2)D$, where D stands for the Relative Mean Deviation:

$$D = \frac{\frac{1}{n} \sum_{i=1}^{n} |y_i - \mu_y|}{\mu_y} \tag{8}$$

Since D is at most equal to $2(n-1)/n$, it follows that these two amounts are bounded by $1-n$ and $n-1$. For the bounded version of the basic level-dependent index, the natural choice for the value of A is therefore $A = 1$.

[2]Strictly speaking, this result holds only when n is even. When n is odd, the sum of the negative weights is $-[(n-1)(n+1)]/(4n)$, which for large n is approximately equal to $-n/4$. And similarly for the positive weights.

2.5 A Brief Comparison

Later in the paper we will consider more complex variants of the indices, but at this stage it seems useful to highlight some of the differences between the basic indices R and L. We focus here on the bounds of the indices, and when these are reached.

The relative version of R varies between $-(n-1)/n$ and $(n-1)/n$, with the minimum value attained when the poorest person in society has all the health, and the maximum when the richest person in society has all the health. It does not matter exactly how poor or how rich the poorest and richest persons are. The relative version of L, by contrast, varies between -1 and $n-1$. The minimum is obtained when the poorest person in society has a zero income and all the health[3]; the maximum when the richest person has all the income and all the health.

If the health variable is unbounded, the minima and maxima of the absolute versions of the indices are equal to those of the relative versions multiplied by the average of the health variable. If the health variable is bounded, the appropriate version of the indices is the bounded one. The minimum and maximum of the bounded version of R, taking $A = 4$, are equal to -1 and 1, with the minimum attained when the poorest half of the population has maximum health and the richest half minimum health, and the maximum in the opposite case. The minimum and maximum of the bounded version of L, with $A = 1$, are equal to $-(1-n)/n$ and $(n-1)/n$. These can be reached only if one person has all the available income; the minimum is attained when that extremely rich person has minimum health, while all the others have maximum health, and the maximum when that person has maximum health, while all the others have minimum health.

3 Decomposition by Health Components

The easiest type of decomposition is by health components. Suppose we have a health variable which can be split up in several parts. A good example is health care expenditures, which may be divided according to source of funding (out-of-pocket costs, health insurance contributions, etc.) or according to type of service (primary care, hospital care, etc.). The starting point is therefore an equation of the following type:

$$h_i = \sum_{j=1}^{m} c_{j,i} \tag{9}$$

[3]More generally, if ties are allowed, the minimum is obtained when all the persons with positive health have zero incomes.

where $c_{j,i}$ stands for the contribution of component c_j $(j = 1, 2, \ldots, m)$ to the health attainment of individual i.

It has been shown that rank-dependent indices can be decomposed by health components without any difficulty (Clarke, Gerdtham and Connelly, 2003), and the same holds for level-dependent indices. It is easy to check that any linear index of type (1) can be written as the sum of the corresponding indices of the health components:

$$I(y,h) = \sum_{j=1}^{m} I(y,c_j) \tag{10}$$

where each component's index $I(y,c_j)$ is defined as:

$$I(y,c_j) = \frac{1}{n}\sum_{i=1}^{n} w_i(y)c_{j,i} \tag{11}$$

The decomposition of the relative index is likewise very simple. The relative index can be expressed as a weighted sum of the relative indices of the components:

$$I_{rel}(y,h) = \sum_{j=1}^{m} \frac{\mu_{c_j}}{\mu_h} I_{rel}(y,c_j) \tag{12}$$

where μ_{c_j} stands for the mean of component c_j and $I_{rel}(y,c_j) = I(y,c_j)/\mu_{c_j}$. The decomposition weights are equal to the shares of the health components in total health. For the bounded index the same formula holds as for the absolute index.

4 Decomposition by Population Groups

4.1 Subgroup Decomposability

The second type of decomposition is by population groups, often called subgroup decomposition. Suppose that we partition the population into k subsets, G_1, G_2, ..., G_k, such that every individual i belongs to exactly one subset G_j. One can think of subgroups based on geographical areas, age, ethnicity, etc. The number of individuals in subgroup G_j is denoted by n_j. The mean income and health attainments of subgroup G_j are equal to:

$$\mu_{y_j} = \frac{1}{n_j}\sum_{i\in G_j} y_i, \quad \mu_{h_j} = \frac{1}{n_j}\sum_{i\in G_j} h_i \tag{13}$$

and obviously we have:

$$\mu_y = \sum_{j=1}^{k} \frac{n_j}{n} \mu_{y_j}, \quad \mu_h = \sum_{j=1}^{k} \frac{n_j}{n} \mu_{h_j} \tag{14}$$

The idea of (additive) subgroup decomposability is to express the measured degree of inequality I as the sum of two parts, the 'within-group' inequality I_W and the 'between-group' inequality I_B (see Bourguignon, 1979). The 'within-group' inequality consists of a weighted sum of the inequalities within the k subgroups $(I_1, I_2, \ldots, I_k)$, with the weights equal to $s_1, s_2, \ldots, s_k$:

$$I_W = \sum_{j=1}^{k} s_j I_j \tag{15}$$

The inequality in subgroup G_j is defined as:

$$I_j = \frac{1}{n_j} \sum_{i \in G_j} w_{i(j)} h_i \tag{16}$$

The notation $w_{i(j)}$ indicates that the weights of individuals depend on their position within each group.

The 'between-group' inequality is calculated assuming that every individual of group G_j has the income level μ_{y_j} and the health level μ_{h_j}. It can therefore be defined as:

$$I_B = \frac{1}{n} \sum_{j=1}^{k} n_j w_{j(B)} \mu_{h_j} \tag{17}$$

The weights $w_{j(B)}$ are applied to every individual of the group G_j and reflect the average situation of an individual of that group.

4.2 Rank-Dependent Indices

From the literature on income inequality we know that univariate rank-dependent indices, such as the Gini coefficient, cannot be decomposed into the sum of a within-group and a between-group contribution, except in exceptional cases (Bhattacharya and Mahalanobis, 1967). The Gini decomposition involves a third or 'residual' component, which some, but not all, find difficult to interpret (see Mookherjee and Shorrocks, 1982, and Lambert and Aronson, 1993, for opposing views on the matter). This result carries over

to rank-dependent bivariate indices (Clarke, Gerdtham and Connelly, 2003; Wagstaff, 2005b).

Let us begin by defining the weights $w_{i(j)}$ and $w_{j(B)}$ of expressions (16) and (17) for the basic rank-dependent index R. For the calculation of the within-group inequalities R_j and their weighted sum $R_W = \sum_{j=1}^{k} s_j R_j$, the relevant ranks are those within each group. If we designate individual i's rank within group G_j, for any $i \in G_j$, by $r_{i(j)}$, the weights $w_{i(j)}$ are equal to:

$$w_{i(j)}^R = \frac{2r_{i(j)} - n_j - 1}{n_j} \tag{18}$$

For the calculation of the between-group inequality R_B, the rank assigned to every individual of a given group must coincide with the average rank of the individuals of that group in the whole population. The average rank $r_{j(B)}$ of the individuals of group G_j can be calculated by means of this recursive formula:

$$r_{j(B)} = \frac{n_j + 1}{2} + \sum_{l=0}^{j-1} n_l \tag{19}$$

with $n_0 = 0$ and $j = 1, \ldots, k$. The weights $w_{j(B)}^R$ are then equal to:

$$w_{j(B)}^R = \frac{2r_{j(B)} - n - 1}{n} \tag{20}$$

It turns out to be impossible to find a set of weights s_j for which we have $R = R_W + R_B$, where the within-group inequality is defined by (15) and the between-group inequality by (17). There is a residual term $R_X = R - R_W - R_B$ which may on occasion be equal to zero, but in general is not. For our calculations, we assume the weights s_j are equal to the population shares n_j/n. In that case, we end up with a residual term equal to:

$$R_X = \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in G_j} \left(w_i^R h_i - w_{i(j)}^R h_i - w_{j(B)}^R \mu_{h_j} \right) \tag{21}$$

Similar results hold for the relative and bounded versions of the index. Hence, and not surprisingly, the basic rank-dependent index does not have the property of additive subgroup decomposability.

4.3 Level-Dependent Indices

The situation is different for the basic level-dependent index. Let us begin by defining the weights to be used for the calculation of the within-group and

between-group inequalities. For the within-group inequalities, the relevant deviations are those of individual i's income from group G_j's mean income, for any $i \in G_j$. For the basic level-dependent index L the weights $w_{i(j)}$ are therefore equal to:

$$w^L_{i(j)} = \frac{y_i - \mu_{y_j}}{\mu_{y_j}} \tag{22}$$

With regard to the calculation of the between-group inequality, each individual of a given group receives a weight proportional to the deviation of the mean income of that group from the mean income of the whole population. This means that the weights $w_{j(B)}$ are equal to:

$$w^L_{j(B)} = \frac{\mu_{y_j} - \mu_y}{\mu_y} \tag{23}$$

Given these definitions and expressions, we can show that the basic level-dependent index L has the property of additive subgroup decomposability.

Theorem 1 *The basic level-dependent index L can be expressed as the sum of a within-group component L_W and a between-group component L_B, where $L_W = \sum_{j=1}^{k} s_j L_j$ and $s_j = (n_j \mu_{y_j})/(n\mu_y)$.*

Proof. The inequality within group G_j, given the weights (22) and after isolation of the term $1/\mu_{y_j}$, turns out to be equal to:

$$L_j = \frac{1}{n_j \mu_{y_j}} \sum_{i \in G_j} (y_i - \mu_{y_j}) h_i$$

Since:

$$\sum_{i \in G_j} (y_i - \mu_{y_j}) h_i = \sum_{i \in G_j} y_i h_i - \mu_{y_j} \sum_{i \in G_j} h_i$$

it follows that:

$$L_j = \frac{1}{n_j \mu_{y_j}} \sum_{i \in G_j} y_i h_i - \mu_{h_j}$$

Multiplying L_j by s_j, and using the fact that $s_j/(n_j\mu_{y_j}) = 1/(n\mu_y)$, we obtain:

$$s_j L_j = \frac{1}{n\mu_y} \sum_{i \in G_j} y_i h_i - s_j \mu_{h_j}$$

Summing over all groups, we find that:

$$\sum_{j=1}^{k} s_j L_j = \frac{1}{n\mu_y} \sum_{j=1}^{k} \sum_{i \in G_j} y_i h_i - \sum_{j=1}^{k} s_j \mu_{h_j}$$

which can be simplified as:

$$\sum_{j=1}^{k} s_j L_j = \frac{1}{n\mu_y} \sum_{i=1}^{n} y_i h_i - \sum_{j=1}^{k} s_j \mu_{h_j}$$

Given (23), the between-group inequality is equal to:

$$L_B = \frac{1}{n\mu_y} \sum_{j=1}^{k} n_j (\mu_{y_j} - \mu_y) \mu_{h_j}$$

Since:

$$\sum_{j=1}^{k} n_j (\mu_{y_j} - \mu_y) \mu_{h_j} = \sum_{j=1}^{k} n_j \mu_{y_j} \mu_{h_j} - \mu_y \sum_{j=1}^{k} n_j \mu_{h_j}$$

it follows that:

$$L_B = \sum_{j=1}^{k} \frac{n_j \mu_{y_j}}{n\mu_y} \mu_{h_j} - \sum_{j=1}^{k} \frac{n_j}{n} \mu_{h_j}$$

which can be simplified as:

$$L_B = \sum_{j=1}^{k} s_j \mu_{h_j} - \mu_h$$

The overall extent of inequality is equal to:

$$L = \frac{1}{n\mu_y} \sum_{i=1}^{n} (y_i - \mu_y) h_i$$

Since:

$$\sum_{i=1}^{n} (y_i - \mu_y) h_i = \sum_{i=1}^{n} y_i h_i - \mu_y \sum_{i=1}^{n} h_i$$

we find that

$$L = \frac{1}{n\mu_y} \sum_{i=1}^{n} y_i h_i - \mu_h$$

Hence, $\sum_{j=1}^{k} s_j L_j + L_B = L$. ■

The decomposition weights of the within-group inequalities are equal to the shares of the groups in total income. Since these shares add up to 1, so do the weights: $\sum_{j=1}^{k} s_j = 1$. That may be perceived as an attractive property.

For the relative version of the index, the decomposition weights are $s_j = (n_j \mu_{y_j} \mu_{h_j})/(n\mu_y \mu_h)$. For the bounded version, the weights are the same as those for the absolute version.

5 Regression-Based Decomposition

Regression-based decomposition methods are now widely used in the analysis of the socioeconomic inequality of health. The approach proposed by Wagstaff, Van Doorslaer and Watanabe (2003) has paved the way for a stream of studies in which the Concentration Index is decomposed based on regressions of health. The basic idea is to start from a simple equation which explains the health variable in terms of the independent variables $x_1, x_2, \ldots, x_q$:

$$h_i = \lambda_0 + \lambda_1 x_{1,i} + \lambda_2 x_{2,i} + \ldots + \lambda_q x_{q,i} + \varepsilon_i \tag{24}$$

where ε_i represents the error term. The right-hand side is then plugged into the formula for the Concentration Index instead of h_i. Applying the procedure to (5), we obtain:

$$R(y,h) = \sum_{j=1}^{q} \hat{\lambda}_j R(y,x_j) + R(y,\varepsilon) \tag{25}$$

Often the terms $\hat{\lambda}_j R(y,x_j)$ are identified as the contributions of the variables x_j to the measured amount of inequality $R(y,h)$, and the term $R(y,\varepsilon)$ as the unexplained part of the inequality.

Obviously, the same procedure can be applied to (7), which then yields:

$$L(y,h) = \sum_{j=1}^{q} \hat{\lambda}_j L(y,x_j) + L(y,\varepsilon) \tag{26}$$

In previous work we have taken a critical look at a few aspects of this approach, and advocated caution in the interpretation of the results (Erreygers and Kessels, 2013; Kessels and Erreygers, 2015). Remaining in the framework of rank-dependent indicators, we also suggested a few alternatives. Here we go a step further and explore a new regression-based method which can be applied to both rank- and level-dependent indicators, but which is probably more suitable for the latter.

Let us begin with the basic (absolute) version of the level-dependent index. Observe that we can write index L as:

$$L = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\mu_y} h_i - \mu_h \right) \tag{27}$$

Defining $z_i = \dfrac{y_i}{\mu_y} h_i - \mu_h$, we arrive at the following extremely simple expression of the index:

$$L = \frac{1}{n} \sum_{i=1}^{n} z_i = \mu_z \tag{28}$$

The value of the z variable can be interpreted as a score which provides an indication of an individual's deviation from the expected outcome in the income-health domain. The mean or expected level of an individual with respect to income is μ_y, and that with respect to health μ_h. Without further information, the expected value of z is therefore $\frac{\mu_y}{\mu_y}\mu_h - \mu_h = 0$. If an individual performs better than expected on both income and health, then $z_i > 0$; in the opposite case, we have $z_i < 0$. If one of the two is higher than expected but the other lower, the sign of z_i is undetermined. If $z_i > 0$, this can be interpreted as indicating an outcome which is on the whole better than expected, and $z_i < 0$ as indicating an outcome which is on the whole worse than expected.

The new decomposition method we propose here is based on a regression of the individual z values. Since the z variable takes into account both income and health, we see this as a more natural procedure to explain the correlation between income and health than the one based on a regression of health only. Assume that we explain z by means of the independent variables $x_1, x_2, \ldots, x_q$. More specifically, let us consider the simple linear regression equation:

$$z_i = \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{2,i} + \ldots + \gamma_q x_{q,i} + \varepsilon_i \tag{29}$$

where ε_i is a well-behaved error term. One certainly expects that variables which are positively associated with both income and health have positive coefficients; and the opposite, of course, in the case of negative association.

Plugging the estimates of (29) into (28), and taking into account that $\mu_{\hat{\varepsilon}} = 0$, we obtain the following result:

$$L = \hat{\gamma}_0 + \hat{\gamma}_1 \mu_{x_1} + \hat{\gamma}_2 \mu_{x_2} + \ldots + \hat{\gamma}_q \mu_{x_q} \tag{30}$$

Expression (30) can be used to assess the expected marginal effect of each independent variable on L. The estimated effect of a small ceteris paribus change $\Delta\mu_{x_j}$ of μ_{x_j} on L is equal to $\hat{\gamma}_j \Delta\mu_{x_j}$. For comparable changes $\Delta\mu_{x_j}$ and $\Delta\mu_{x_l}$ of the variables x_j and x_l, it is therefore possible to determine which of two has the largest marginal effect on the index. What we definitely should not do, is claim that the product $\hat{\gamma}_j \mu_{x_j}$ constitutes the contribution of variable x_j to the observed level of inequality. If that were true, the contribution of any mean centered variable would automatically be zero.

Can the same approach also be applied to rank-dependent indicators? It can, with one important change. Let us start by observing that the basic rank-dependent index R may be expressed as:

$$R = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{r_i - \frac{1}{2}}{\frac{n}{2}} \right) h_i - \mu_h \right] \tag{31}$$

The mean rank of an individual is equal to $(n+1)/2$. If we expect individual i to attain that rank, the expected value of $r_i - 1/2$ is equal to $n/2$. It follows that the expected value of the expression between square brackets in (31) is 0. Designating this expression as u_i, we can interpret $u_i > 0$ as an indication that individual i attains an outcome in the income rank-health space which is on the whole better than expected, and $u_i < 0$ as indicating an outcome which is on the whole worse than expected.

Since we now have:

$$R = \frac{1}{n}\sum_{i=1}^{n} u_i = \mu_u \tag{32}$$

what we should do next is to explain u by means of a set of independent variables. Suppose that we estimate the following regression:

$$u_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_q x_{q,i} + \eta_i \tag{33}$$

We then end up with the following counterpart of (30):

$$R = \hat{\beta}_0 + \hat{\beta}_1 \mu_{x_1} + \hat{\beta}_2 \mu_{x_2} + \ldots + \hat{\beta}_q \mu_{x_q} \tag{34}$$

An important difference between the decomposition of L and that of R, is that the expected situation of individual i in the first case corresponds to the mean income μ_y and the mean health μ_h, but in the second to the mean income rank (and hence to the *median* rather than the mean income) and the mean health.

6 Extended Indices

In this section we examine to what extent the decomposition results for the basic versions of the rank-dependent and level-dependent indices carry over to the so-called extended versions of these indices.[4] These extended versions are similar to the distribution-sensitive indices developed in the income inequality literature (see, e.g. Yitzhaki, 1983; Lambert and Lanza, 2006).

6.1 Defining Extended Indices

The basic rank-dependent index is characterized by a weighting function which is linear in the income ranks, viz. (4). Other versions, such as the extended Concentration Index (Pereira, 1998; Wagstaff, 2002), have been

[4]Since our empirical application is limited to the basic versions, this section can be omitted without loss of continuity by readers who are less interested in technicalities.

developed to give relatively more weight to individuals with lower ranks. The extended version of the rank weighting function can be expressed as follows:

$$w_i^R(\nu) = \frac{1 + n\left[\left(\frac{n-r_i}{n}\right)^\nu - \left(\frac{n-r_i+1}{n}\right)^\nu\right]}{\nu - 1} \tag{35}$$

where $\nu \geq 2$ is a distributional sensitivity parameter (see, e.g., Erreygers, Clarke and Van Ourti, 2012). The extended rank-dependent index is denoted by $R(\nu)$:

$$R(\nu) = \frac{1}{n}\sum_{i=1}^{n} w_i^R(\nu) h_i \tag{36}$$

The linear weighting function (4) corresponds to the value $\nu = 2$. The higher the value of ν, the more sensitive the index is to changes in the lower end of the rank distribution. The sum of the negative weights is equal to $-n/(\nu^{\nu/(\nu-1)})$, and that of the positive weights $n/(\nu^{\nu/(\nu-1)})$. For $\nu > 2$, the lowest negative weight is (approximately) -1, and the highest positive weight (approximately) $1/(\nu - 1)$.

It is also possible to make the level-dependent index more sensitive to the bottom of the income distribution. One way of doing this consists of transforming the incomes. Erreygers and Kessels (2015) propose a transformation based on the following isoelastic function:

$$y_i(\alpha) = \begin{cases} \frac{y_i^{1-\alpha} - \alpha}{1-\alpha} & (\alpha \neq 1) \\ 1 + \log(y_i) & (\alpha = 1) \end{cases} \tag{37}$$

but other functions may be used as well.[5] The mean of this transformed income is defined as $\mu_{y(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} y_i(\alpha)$. Erreygers and Kessels (2015) suggest the following generalization of the linear weighting function (6) which characterizes the basic level-dependent index:

$$w_i^L(\alpha) = \frac{y_i(\alpha) - \mu_{y(\alpha)}}{\sum_{j=1}^{n}\left|y_j(\alpha) - \mu_{y(\alpha)}\right|} . \frac{\sum_{j=1}^{n}|y_j - \mu_y|}{\mu_y} \tag{38}$$

The corresponding extended level-dependent index is denoted by $L(\alpha)$:

$$L(\alpha) = \frac{1}{n}\sum_{i=1}^{n} w_i^L(\alpha) h_i \tag{39}$$

[5] A disadvantage of using function (37) is that $y_i(\alpha)$ no longer exists when $y_i = 0$ and $\alpha \geq 1$. This excludes the use of data according to which some individuals have zero or negative incomes, as is often the case. Instead of income, one might consider using consumption data.

An alternative way of writing the weighting function makes it explicit how exactly the transformation of incomes affects the weights. Given that D stands for the Relative Mean Deviation of y, let us define $D(\alpha)$ as the Relative Mean Deviation of $y(\alpha)$, i.e. $D(\alpha) = \left(\sum_{j=1}^{n} \left|y_j(\alpha) - \mu_{y(\alpha)}\right|\right) / \mu_{y(\alpha)}$. We designate the ratio of $D(\alpha)$ to D by $\phi(\alpha)$:

$$\phi(\alpha) = \frac{D(\alpha)}{D} \tag{40}$$

Since the transformation has a greater effect on high incomes than on low incomes, $\phi(\alpha)$ measures to what extent income inequality decreases as a result of the transformation. The ratio is used to 'inflate' the weights. As a matter of fact, the weights (38) can be written as:

$$w_i^L(\alpha) = \frac{y_i(\alpha) - \mu_{y(\alpha)}}{\mu_{y(\alpha)}} . \frac{1}{\phi(\alpha)} \tag{41}$$

Without the inclusion of $1/\phi(\alpha)$, most of the weights would become very small as the value of α increases.

It is easy to see that $\alpha = 0$ corresponds to the basic case: no transformation is applied to the income levels. As α increases, the weight of the most well-off individual remains positive but decreases in magnitude, whereas the weight of the least well-off individual remains negative but increases in magnitude. What happens to the weights of the other individuals depends on the specific income distribution. More and more individuals who initially had a negative weight will get a positive weight, until eventually only the least well-off individual will be the only one with a negative weight. As in the basic case, all the negative weights sum to $-(n/2)D$, and all the positive weights to $(n/2)D$. Since D is at most equal to $2(n-1)/n$, it follows that these two amounts are bounded by $1-n$ and $n-1$. For high values of α, therefore, the lowest negative weight will tend to $-(n/2)D \geq 1-n$, while all the other weights will be positive and tend to $((n/2)D)/(n-1) \leq 1$.

6.2 Decomposition by Population Groups

For the extended rank-dependent index $R(\nu)$, the counterparts of expressions (18) and (20) are

$$w_{i(j)}^R(\nu) = \frac{1 + n\left[\left(\frac{n_j - r_{i(j)}}{n_j}\right)^{\nu} - \left(\frac{n_j - r_{i(j)}+1}{n_j}\right)^{\nu}\right]}{\nu - 1} \tag{42}$$

$$w_{j(B)}^R(\nu) = \frac{1 + n\left[\left(\frac{n - r_{j(B)}}{n}\right)^{\nu} - \left(\frac{n - r_{j(B)}+1}{n}\right)^{\nu}\right]}{\nu - 1} \tag{43}$$

As in the basic case, the index $R(\nu)$ is not subgroup decomposable.

For the extended level-dependent index $L(\alpha)$, the counterpart of expression (22) is:

$$w_{i(j)}^{L}(\alpha) = \frac{y_i(\alpha) - \mu_{y_j(\alpha)}}{\mu_{y_j(\alpha)}} . \frac{1}{\phi_j(\alpha)} \tag{44}$$

where:

$$\mu_{y_j(\alpha)} = \frac{1}{n_j} \sum_{i \in G_j} y_i(\alpha) \tag{45}$$

and:

$$\phi_j(\alpha) = \frac{D_j(\alpha)}{D_j} \tag{46}$$

Likewise, the counterpart of (23) is:

$$w_{j(B)}^{L}(\alpha) = \frac{\mu_{y_j(\alpha)} - \mu_{y(\alpha)}}{\mu_{y(\alpha)}} . \frac{1}{\phi_B(\alpha)} \tag{47}$$

where[6]:

$$\phi_B(\alpha) = \frac{D_B(\alpha)}{D_B} \tag{48}$$

Unfortunately, the property of subgroup decomposability no longer holds for the extended level-dependent index. It turns out to be impossible to obtain a 'clean' subgroup decomposition formula of the type $L(\alpha) = \sum_{j=1}^{k} s_j L_j(\alpha) + L_B(\alpha)$. We can, however, generate an expression which comes close to the desired result.

Theorem 2 *The extended level-dependent index $L(\alpha)$ can be expressed as the sum of a modified within-group component $M_W(\alpha)$ and a modified between-group component $M_B(\alpha)$, where $M_W(\alpha) = \sum_{j=1}^{k} \frac{\phi_j(\alpha)}{\phi(\alpha)} s_j(\alpha) L_j(\alpha)$, $M_B(\alpha) = \frac{\phi_B(\alpha)}{\phi(\alpha)} L_B(\alpha)$, and $s_j(\alpha) = \frac{n_j \mu_{y_j(\alpha)}}{n \mu_{y(\alpha)}}$.*

Proof. The proof proceeds along the same lines as the proof of Theorem 1. The first step consists of showing that:

$$L_j(\alpha) = \frac{1}{\phi_j(\alpha)} \left[\frac{1}{n_j \mu_{y_j(\alpha)}} \sum_{i \in G_j} y_i(\alpha) h_i - \mu_{h_j} \right]$$

[6]We have: $D_B = \left(\frac{1}{n} \sum_{j=1}^{k} n_j \left| \mu_{y_j} - \mu_y \right| \right) / \mu_y$, etc.

and then using this result to establish that:

$$M_W(\alpha) = \sum_{j=1}^{k} \frac{\phi_j(\alpha)}{\phi(\alpha)} s_j(\alpha) L_j(\alpha) = \frac{1}{\phi(\alpha)} \left[\frac{1}{n\mu_{y(\alpha)}} \sum_{i=1}^{n} y_i(\alpha) h_i - \sum_{j=1}^{k} s_j(\alpha) \mu_{h_j} \right]$$

The second step consists of showing that:

$$L_B(\alpha) = \frac{1}{\phi_B(\alpha)} \left[\sum_{j=1}^{k} s_j(\alpha) \mu_{h_j} - \mu_h \right]$$

which implies that:

$$M_B(\alpha) = \frac{\phi_B(\alpha)}{\phi(\alpha)} L_B(\alpha) = \frac{1}{\phi(\alpha)} \left[\sum_{j=1}^{k} s_j(\alpha) \mu_{h_j} - \mu_h \right]$$

The third step consists of showing that:

$$L(\alpha) = \frac{1}{\phi(\alpha)} \left[\frac{1}{n\mu_{y(\alpha)}} \sum_{i=1}^{n} y_i(\alpha) h_i - \mu_h \right]$$

Combining these results, we arrive at the conclusion that $M_W(\alpha) + M_B(\alpha) = L(\alpha)$. ■

Observe that for $\alpha = 0$, we have $\phi(0) = \phi_j(0) = \phi_B(0) = 1$, $\mu_{y(\alpha)} = \mu_y$ and $\mu_{y_j(\alpha)} = \mu_{y_j}$, and therefore $M_W(0) = L_W$ and $M_B(0) = L_B$. For values of α exceeding 0, a distortion occurs which requires the introduction of a modification factor. This is equal to $\phi_j(\alpha)/\phi(\alpha)$ for the inequality within group G_j, and to $\phi_B(\alpha)/\phi(\alpha)$ for the between-group inequality. As long as the $\phi_j(\alpha)$'s and $\phi_B(\alpha)$ do not deviate far from $\phi(\alpha)$, the distortion is small and the modified decomposition formula remains close to the clean result of the basic case. In other words, if the transformation of incomes brings about similar changes in inequality within all groups and between the groups as it does for society as a whole, then the modifications are limited.

To conclude the discussion of subgroup decomposability, we observe that Theorem 2 refers to the absolute version of the extended level-dependent index. For the relative version, the decomposition weights $s_j(\alpha)$ must be multiplied by μ_{h_j}/μ_h.

6.3 Regression-Based Decomposition

When it comes to regression-based decomposition, the procedures to be followed for the extended indices $L(\alpha)$ and $R(\nu)$ are similar to the ones for L

and R. Given (39) and (41), we can express $L(\alpha)$ as:

$$L(\alpha) = \frac{1}{\phi(\alpha)} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i(\alpha)}{\mu_{y(\alpha)}} h_i - \mu_h \right) \tag{49}$$

Defining $z_i(\alpha) = \dfrac{y_i(\alpha)}{\mu_{y(\alpha)}} h_i - \mu_h$, it follows that we can write:

$$L(\alpha) = \frac{1}{\phi(\alpha)} \frac{1}{n} \sum_{i=1}^{n} z_i(\alpha) = \frac{\mu_{z(\alpha)}}{\phi(\alpha)} \tag{50}$$

Assume now that we explain $z(\alpha)/\phi(\alpha)$ by means of the independent variables $x_1, x_2, \ldots, x_q$. More specifically, let us consider the regression equation:

$$\frac{z_i(\alpha)}{\phi(\alpha)} = \gamma_0(\alpha) + \gamma_1(\alpha) x_{1,i} + \gamma_2(\alpha) x_{2,i} + \ldots + \gamma_q(\alpha) x_{q,i} + \varepsilon_i(\alpha) \tag{51}$$

where $\varepsilon_i(\alpha)$ is a well-behaved error term. This leads to the following estimate:

$$L(\alpha) = \hat{\gamma}_0(\alpha) + \hat{\gamma}_1(\alpha) \mu_{x_1} + \hat{\gamma}_2(\alpha) \mu_{x_2} + \ldots + \hat{\gamma}_q(\alpha) \mu_{x_q} \tag{52}$$

The expected effect of a small ceteris paribus change $\Delta\mu_{x_j}$ of μ_{x_j} on $L(\alpha)$ is equal to $\hat{\gamma}_j(\alpha) \Delta\mu_{x_j}$.

For the relative index $L_{rel}(\alpha) = L(\alpha)/\mu_h$, the same regression can be used. All that needs to be changed is that the expected effect of a small change $\Delta\mu_{x_j}$ of μ_{x_j} on $L_{rel}(\alpha)$ is equal to $\left[\hat{\gamma}_j(\alpha) \Delta\mu_{x_j}\right] / \mu_h$.

For the extended rank-dependent index we have to write the index as:

$$R(\nu) = \frac{1}{n} \sum_{i=1}^{n} u_i(\nu) \tag{53}$$

where $u_i(\nu)$ is defined as:

$$u_i(\nu) = \left(\frac{\nu + n \left[\left(\frac{n - r_i}{n} \right)^{\nu} - \left(\frac{n - r_i + 1}{n} \right)^{\nu} \right]}{\nu - 1} \right) h_i - \mu_h \tag{54}$$

The corresponding regression equation is then:

$$u_i(\nu) = \beta_0(\nu) + \beta_1(\nu) x_{1,i} + \beta_2(\nu) x_{2,i} + \ldots + \beta_q(\nu) x_{q,i} + \eta_i(\nu) \tag{55}$$

which leads to the decomposition:

$$R(\nu) = \widehat{\beta}_0(\nu) + \widehat{\beta}_1(\nu) \mu_{x_1} + \widehat{\beta}_2(\nu) \mu_{x_2} + \ldots + \widehat{\beta}_q(\nu) \mu_{x_q} \tag{56}$$

While it is relatively straightforward to give meaning to the sign of the variable $z_i(\alpha)$ used in the decomposition of the level-dependent index $L(\alpha)$, it seems more difficult to do so for that of the variable $u_i(\nu)$ used in the decomposition of the rank-dependent index $R(\nu)$.

7 Empirical Application

7.1 Description of the Data

For the empirical illustration we limit ourselves to the basic rank- and level-dependent indices R and L, and we look at the subgroup and regression-based decompositions only.

Our data come from the Household, Income and Labour Dynamics in Australia (HILDA) survey, wave 13.[7] As our income variable we take equivalised income, calculated using the modified OECD equivalence scale (Australian Bureau of Statistics, 2013, Appendix 3). Each member of a household is assigned an equivalised income equal to the household's disposable income divided by its equivalence factor. This factor is obtained by giving the first adult of the household a score of 1, every other adult a score of 0.5, every child a score of 0.3, and then adding up all the scores. As our health variable we take the SF-6D health score. Since this is a bounded variable, we use the bounded version of the inequality indices, i.e. R_{bou} and L_{bou}. Given that the bounds of the variable are $h_{\min} = 0$ and $h_{\max} = 1$, it is easy to check that $R_{bou} = 4R$ and $L_{bou} = L$.

For the subgroup decomposition we consider partitions based on sex, age and employment status. For the regression-based decomposition we additionally take into account variables measuring individuals' education level, housing type, number of children, marital status, physical activity, and whether they are suffering from any long-term health conditions. Our sample includes $14,729$ individuals with valid, non-missing values for all variables under study.[8] Since this is a fairly large and, to the best of our knowledge, representative sample, we decided not to apply sample weights in our calculations. The composition of our sample is described in Table 1.

[7] More information on the HILDA database can be found here: `https://www.melbourneinstitute.com/hilda/`.

[8] This means in particular that individuals with a negative income are excluded.

Table 1: Frequency statistics of the sample

Variable (in % unless stated otherwise)			
Sex		Marital status	
Female	53.17	Married	64.18
Male	46.83	Not married	35.72
Age group		Housing type	
15-24	17.43	Non private dwelling	0.24
25-34	16.97	House	81.28
35-44	16.00	Semi-attached house	6.74
45-54	17.66	Flat/apartment	11.24
55-64	14.64	Other dwelling type	0.50
65-74	10.65		
75+	6.63		
Employment status		Education level	
Employed	63.49	Less than secondary education	27.99
Unemployed	4.04	Secondary education	37.33
Not in labour force	32.47	Higher education	34.68
Number of children		Having a long-term health condition	
Aged 0-4 (average)	0.18	Yes	29.64
Aged 5-14 (average)	0.30	No	70.36
Physical activity			
No	11.11		
Some	37.27		
Frequent	51.62		

7.2 Decomposition by Population Groups

For the first subgroup decomposition we partitioned the population into two groups, females and males. As shown in Table 2, there is some variation among these groups, with respect to both health and income. Women tend to have slightly worse health and lower income than men. The rank-dependent index measures a slightly higher extent of inequality among women than among men, while the level-dependent index detects no difference between the two groups.

Table 2: Summary statistics for the sex decomposition

	Health		Income ($)		Indices	
Group	Mean	S.D.	Mean	S.D.	R_j	L_j
Female	0.7508	0.1244	50,014	34,725	0.0661	0.0141
Male	0.7718	0.1216	52,389	35,102	0.0629	0.0141
All	0.7606	0.1236	51,126	34,921	0.0657	0.0144

The results of the subgroup decomposition according to sex can be found in Table 3. Let us begin by looking at the results for the level-dependent index. The observed inequality comes overwhelmingly from heterogeneity within the groups; the between-component accounts for a meagre 1.70% of the total. The results for the rank-dependent index also point in the direction of the predominance of the within-group component, but suggest a much higher contribution of the between-component, equal to 31.92% of the total. These results must be treated with caution, however, since the residual component amounts to roughly −30%.

Table 3: Decomposition according to sex

	R		L	
	Values	%	Values	%
Within	0.0646	98.39%	0.0141	98.30%
Between	0.0210	31.92%	0.0002	1.70%
Residual	−0.0199	−30.31%	–	–
Total	0.0657	100.00%	0.0144	100.00%

For our second subgroup decomposition we divided the population into seven age groups, with the youngest group consisting of those aged 15 to 24, and the oldest group consisting of those aged 75 or older. As can be seen from Table 4, older people tend to have worse health and lower income than younger people. According to both indices, the level of socioeconomic inequality is the highest in the group aged 55 to 64.

Table 4: Summary statistics for the age decomposition

	Health		Income ($)		Indices	
Group	Mean	S.D.	Mean	S.D.	R_j	L_j
Age 15-24	0.7837	0.1179	45,376	26,065	0.0334	0.0080
Age 25-34	0.7841	0.1132	55,126	29,572	0.0490	0.0106
Age 35-44	0.7741	0.1204	54,101	31,592	0.0555	0.0114
Age 45-54	0.7586	0.1218	56,699	33,753	0.0654	0.0130
Age 55-64	0.7438	0.1266	55,639	40,009	0.0818	0.0192
Age 65-74	0.7333	0.1270	45,305	46,554	0.0581	0.0167
Age 75+	0.6937	0.1249	33,381	35,607	0.0330	0.0110
All	0.7606	0.1236	51,126	34,921	0.0657	0.0144

The decomposition results are reported in Table 5. According to both indices the within-component is again by far the largest. In comparison to the results for the decomposition by sex, the level-dependent index now indicates a substantially higher between-component (11.14% instead of 1.70%), but the rank-dependent index a substantially lower one (20.17% instead of 31.92%). This is a striking difference.[9]

[9]We also calculated the within- and between-components for a combined sex and age subgroup classification, i.e. based on 14 sex-age groups. The level-dependent index estimates that the between-component amounts to 12.48%, which is close to the sum of 1.70% and 11.14%; the rank-dependent index, by contrast, estimates that this is equal to 37.12%, which is nowhere near the sum of 31.90% and 20.17%.

Table 5: Decomposition according to age

	R		L	
	Values	%	Values	%
Within	0.0549	83.59%	0.0126	88.86%
Between	0.0132	20.17%	0.0018	11.14%
Residual	−0.0025	−3.75%	–	–
Total	0.0657	100.00%	0.0144	100.00%

For our third example of subgroup decomposition we partitioned the population in three groups according to their employment status. Table 6 indicates that there are large differences in health and income between the employed, the unemployed and those who are not in the labour force, and also that the level of inequality is the highest among the last.

Table 6: Summary statistics for the employment decomposition

	Health		Income ($)		Indices	
Group	Mean	S.D.	Mean	S.D.	R_j	L_j
Employed	0.7857	0.1088	58,553	34,503	0.0213	0.0051
Unemployed	0.7406	0.1313	37,811	23,340	0.0353	0.0096
Not in labour force	0.7140	0.1352	38,263	32,638	0.0645	0.0165
All	0.7606	0.1236	51,126	34,921	0.0657	0.0144

Table 7 gives the decomposition results for both indices. The level-dependent index attributes 44.11% of the observed inequality to the between-group component, which is much larger than what we obtained for the two previous decompositions. The rank-dependent index, by contrast, suggests that the between-component is much larger than the within-component, and moreover that it is almost equal in size to the overall inequality. Once again, given the huge residual term (−49.59% of the observed inequality!) this result must be treated with caution.

Table 7: Decomposition according to employment

	R		L	
	Values	%	Values	%
Within	0.0359	54.63%	0.0080	55.89%
Between	0.0624	94.96%	0.0063	44.11%
Residual	−0.0326	−49.59%	–	–
Total	0.0657	100.00%	0.0144	100.00%

From these examples we conclude three things. First, for both indices we always obtain roughly the same estimates of the share of the within-group component. Second, with regard to the share of the between-group component, the estimates for the rank-dependent index tend to be much higher than those for the level-dependent index. And third, the presence of a substantial residual term complicates the interpretation of the results for the rank-dependent index.

7.3 Regression-based Decomposition

As far as the regression-based decomposition is concerned, we compare the health-oriented approach with the alternative approach suggested in the paper. We would like to stress, however, that while the first is often used to identify the relative contribution of different factors to the socioeconomic inequality of health, the second approach only allows us to estimate the marginal effect of these factors on the observed level of inequality.

Table 8: Summary statistics for the additional variables

	Health		Income ($)		Indices	
Variable	Mean	S.D.	Mean	S.D.	R_j	L_j
Education						
Less than secondary	0.7368	0.1317	40,048	31,833	0.0773	0.0165
Secondary	0.7630	0.1237	48,423	30,288	0.0550	0.0115
Higher	0.7773	0.1132	62,976	38,266	0.0438	0.0093
Housing type						
Non-private dwelling	0.6772	0.1408	21,399	14,107	−0.0356	−0.0092
House	0.7625	0.1225	51,257	33,281	0.0586	0.0124
Semi-detached house	0.7523	0.1260	51,958	41,341	0.0983	0.0226
Flat/apartment	0.7553	0.1281	50,966	41,969	0.0865	0.0219
Other dwelling type	0.7280	0.1253	36,750	23,051	0.0942	0.0197
Number of children*						
Aged 0-4	0.7821	0.1145	47,050	26,270	0.0449	0.0091
Aged 5-14	0.7679	0.1210	48,811	29,359	0.0648	0.0124
Marital status						
Married	0.7658	0.1195	55,058	35,959	0.0558	0.0118
Not married	0.7513	0.1300	44,082	31,790	0.0771	0.0178
Physical activity						
No	0.6752	0.1416	42,845	33,565	0.0894	0.0210
Some	0.7505	0.1185	50,740	34,244	0.0581	0.0127
Frequent	0.7863	0.1132	53,187	35,420	0.0473	0.0105
Long-term condition						
Yes	0.6753	0.1276	42,892	33,169	0.0695	0.0181
No	0.7966	0.1024	54,595	35,061	0.0239	0.0054
All	0.7606	0.1236	51,126	34,921	0.0657	0.0144

* Averages for the number of individuals with children aged 0-4 (1976) and with children aged 5-14 (2789).

We consider three regressions, and each of these uses the same set of independent variables. The dependent variable of our first regression is health (h), that of our second regression the combined income rank-health variable (u), and that of our third regression the combined income-health variable (z). The independent variables are those mentioned in Table 1. Table 8 provides summary statistics for the variables other than sex, age and employment status. Except for the number of children aged 0-4 and aged 5-15, all variables enter into the regression equation as dummy variables. The reference situation is that of a female person, aged 15-24, who has completed higher education, is employed, lives in a house, is married, does frequent physical activity, and does not suffer from a long-term health condition.

The results of the first regression, and the associated decompositions of the R and L indices, can be found in Table 9. The regression coefficients $\widehat{\lambda}_j$ suggest that being male and having more young children have positive effects on health; all other variables either have negative effects, or are insignificant. The contribution of variable x_j to R and L depends on the estimated coefficient $\widehat{\lambda}_j$ and on the indices $R(y, x_j)$ and $L(y, x_j)$.[10] The results for index R are quite similar to those for index L. The largest contribution seems to come from having a long-term health condition, followed by not being in the labour force and by not doing any physical activity. Not surprisingly, these are factors which are both negatively associated with health (as can be seen from the negative sign of the regression coefficients $\widehat{\lambda}_j$) and characterized by a strong pro-poor distribution (as revealed by the indices $R(y, x_j)$ and $L(y, x_j)$). Most contributions are positive, but there are also several factors which contribute negatively, such as being male and having a higher number of young children. It must also be observed that the so-called unexplained part of the decomposition is quite substantial, viz. 26.19% (= 0.0172/0.0657) for the R index and 31.24% (= 0.0045/0.0144) for the L index.

[10]It may be useful to point out that the indices $R(y, x_j)$ and $L(y, x_j)$ differ from the indices R_j and L_j reported in the summary statistics tables. While the former are calculated over the *whole* population, the latter are *subgroup* indices (e.g., limited to the male population).

Table 9: The health regression and the decompositions of R and L

Variable	$\widehat{\lambda}_j$	S.E.	$R(y, x_j)$	$\widehat{\lambda}_j R(y, x_j)$	$L(y, x_j)$	$\widehat{\lambda}_j L(y, x_j)$
Male	0.0118	(0.0018)	0.0539	0.0006	0.0116	0.0001
Age 25-34	−0.0127	(0.0033)	0.0830	−0.0011	0.0133	−0.0002
Age 35-44	−0.0143	(0.0037)	0.0571	−0.0008	0.0093	−0.0001
Age 45-54	−0.0204	(0.0033)	0.0966	−0.0020	0.0192	−0.0004
Age 55-64	−0.0162	(0.0034)	0.0403	−0.0007	0.0129	−0.0002
Age 65-74	0.0007°	(0.0039)	−0.0920	−0.0001	−0.0121	0.0000
Age 75+	−0.0039°	(0.0045)	−0.1241	0.0005	−0.0230	0.0001
Less than second. edu.	−0.0090	(0.0024)	−0.2919	0.0026	−0.0606	0.0005
Secondary education	−0.0083	(0.0021)	−0.0378	0.0003	−0.0197	0.0002
Unemployed	−0.0328	(0.0049)	−0.0464	0.0015	−0.0105	0.0003
Not in labour force	−0.0357	(0.0024)	−0.4244	0.0152	−0.0817	0.0029
Non-private dwelling	−0.0300§	(0.0178)	−0.0067	0.0002	−0.0014	0.0000
Semi-detached house	−0.0094	(0.0036)	−0.0054	0.0001	0.0011	0.0000
Flat	−0.0087	(0.0029)	−0.0203	0.0002	−0.0004	0.0000
Other dwelling	−0.0162°	(0.0126)	−0.0066	0.0001	−0.0014	0.0000
Children 0-4	0.0061	(0.0019)	−0.0641	−0.0004	−0.0177	−0.0001
Children 5-14	−0.0023°	(0.0015)	−0.0625	0.0001	−0.0164	0.0000
Single	−0.0106	(0.0021)	−0.2148	0.0023	−0.0494	0.0005
No physical activity	−0.0714	(0.0033)	−0.0888	0.0063	−0.0180	0.0013
Some physical activity	−0.0279	(0.0019)	0.0011	0.0000	−0.0028	0.0001
Long-term health cond.	−0.0988	(0.0023)	−0.2372	0.0234	−0.0477	0.0047
Residual term			0.0172	0.0172	0.0045	0.0045
R^2	0.2632					

Bootstrap standard errors between parentheses.

§ = significant at the 10% level, o = insignificant. All other coefficients are significant at the 1% level.

The results for the regression of the composite income rank-health variable u and of the composite income-health variable z are reported in Table 10. Being male and not too old tends to have a positive effect on an individual's achievement in the income rank-health and income-health domains; all the other variables either have negative coefficients or are insignificant. Whereas the health regression showed that having more young children is positively associated with health, the two composite variable regressions reveal that the opposite holds with regard to the income rank-health and income-health achievements. We can also compare the magnitudes of the marginal effects, with the exception of those of the two variables capturing the number of children aged 0 to 4 and aged 5 to 14. As a matter of fact, since all other variables are dummy variables, a given change in their means has a uniform meaning. For instance, an increase of 0.01 of the mean value of the male dummy stands for a 1 percentage point increase in the prevalence of men in the population (of course offset by a 1 percentage point decrease in the prevalence of women). We can therefore say that as far as the R index is concerned, the predicted marginal effect is the highest for a change in the prevalence of the unemployed, followed by changes in the prevalence of those who live in non-private dwellings, who are not in the labour force, and who have less than secondary education. As far as the L index is concerned, the magnitude of the predicted marginal effects is the largest for those who have less than secondary education, followed by those who are unemployed, who are not in the labour force, and who live in a non-private dwelling. It deserves to be pointed out that the effects of these variables are all negative; this means that an increase in the prevalence of the unemployed, to give just one example, is associated with a pro-poor change of the indices.

Table 10: Regression results for the regressions of u and z

Variable	$\widehat{\beta}_j$	S.E.	$\widehat{\gamma}_j$	S.E.
Male	0.0930	(0.0266)	0.0254	(0.0086)
Age 25-34	0.2327	(0.0511)	0.0422	(0.0144)
Age 35-44	0.2510	(0.0554)	0.0519	(0.0152)
Age 45-54	$0.1142^{\#}$	(0.0513)	$0.0353^{\#}$	(0.0141)
Age 55-64	0.0246°	(0.0531)	0.0478	(0.0168)
Age 65-74	−0.2785	(0.0615)	0.0142°	(0.0248)
Age 75+	−0.5799	(0.0637)	−0.0681	(0.0222)
Less than secondary education	−0.9917	(0.0355)	−0.2411	(0.0120)
Secondary education	−0.7012	(0.0312)	−0.1941	(0.0108)
Unemployed	−1.1330	(0.0710)	−0.2341	(0.0163)
Not in labour force	−1.0612	(0.0381)	−0.2223	(0.0135)
Non-private dwelling	−1.0678	(0.2506)	−0.2175	(0.0495)
Semi-detached house	−0.1702	(0.0537)	-0.0059°	(0.0194)
Flat	−0.3053	(0.0436)	$-0.0296^{\S}$	(0.0170)
Other dwelling	−0.7373	(0.1822)	−0.1593	(0.0373)
Children 0-4	−0.4769	(0.0270)	−0.1004	(0.0092)
Children 5-14	−0.3659	(0.0210)	−0.0736	(0.0069)
Single	−0.5500	(0.0301)	−0.1243	(0.0099)
No physical activity	−0.4259	(0.0420)	−0.1098	(0.0137)
Some physical activity	−0.1787	(0.0279)	−0.0505	(0.0094)
Long-term health condition	−0.6706	(0.0320)	−0.1653	(0.0100)
Constant	1.6763	(0.0506)	0.3836	(0.0150)
R^2	0.3199		0.1748	

Bootstrap standard errors between parentheses.

= significant at the 5% level, § = significant at the 10% level, o = insignificant. All other coefficients are significant at the 1% level.

From these results it is clear that the health-oriented decomposition proposed by Wagstaff, Van Doorslaer and Watanabe is of a different nature than the two composite variable oriented decompositions suggested above. While the former identifies risk factors such as not being in the labour force and suffering from a long-term health condition as contributing positively to the pro-rich social gradient of well-being, the latter sees increases in these variables as having a pro-poor effect on the social gradient. This illustrates that great caution must be exercised in the interpretation of regression-based decomposition results.

8 Conclusion

In this paper we explored the decomposition properties of both rank-dependent and level-dependent indices of socioeconomic inequality of health. As far as decomposition by components and regression-based decomposition is concerned, there are no essential differences between the types of indices. When it comes to decomposition by population subgroups, however, level-dependent indices are clearly superior. The fact that the basic level-dependent index can be decomposed perfectly into a 'within' and a 'between' component, and the extended level-dependent index nearly so, constitutes a strong argument in favour of using these indices alongside, and maybe even instead of, the still dominant rank-dependent indices.

Appendix: Sample Weights and Ties

This appendix describes how the weights w_i must be defined when working with datasets in which not all individuals have the same sample weight, and when there are ties between individuals in the income distribution.

Let us assume that the individuals of our dataset $(1, 2, \ldots, n)$ are ranked according to their individual income, i.e. $y_1 \leq y_2 \leq \ldots \leq y_n$. We denote the sample weight of individual i by σ_i, and we assume that $\sum_{i=1}^{n} \sigma_i = 1$.

If there are ties in the income distribution, we define k groups of individuals $G_1, G_2, \ldots, G_k$ such that everyone in group G_j has income y_{G_j}, and moreover $y_{G_1} < y_{G_2} < \ldots < y_{G_k}$. The sample weight of group G_j is $\sigma_{G_j} = \sum_{i \in G_j} \sigma_i$. The cumulative sample weight of group G_j is defined by the recursive formula $\pi_{G_j} = \pi_{G_{j-1}} + \sigma_{G_j}$, where we take $\pi_{G_0} = 0$ and $j = 1, \ldots, k$.

Rank-Dependent Weights

Let individual i belong to group G_j. Then for $\nu \geq 2$ the weight of this individual is equal to:

$$w_i^R(\nu) = \left(\frac{1}{\nu - 1}\right) n \left(\sigma_i + \left(\frac{\sigma_i}{\sigma_{G_j}}\right) \left[\left(1 - \pi_{G_j}\right)^\nu - \left(1 - \pi_{G_{j-1}}\right)^\nu\right]\right) \quad \text{(A.1)}$$

which for the basic version ($\nu = 2$) reduces to:

$$w_i^R = n\sigma_i(\pi_{G_{j-1}} + \pi_{G_j} - 1) \quad \text{(A.2)}$$

Working out the terms between brackets, we obtain:

$$w_i^R = n\sigma_i \left[\sigma_{G_j} + 2\sum_{l=0}^{j-1} \sigma_{G_l} - 1\right] \quad \text{(A.3)}$$

Level-Dependent Weights

When working with level-dependent weights, there is no need to consider group weights. We do have to modify the definitions of the mean income $\mu_{y(\alpha)}$ and of the Relative Mean Deviation $D(\alpha)$. The weighted mean income is now $\mu_{y(\alpha)} = \sum_{i=1}^n \sigma_i y_i(\alpha)$, and the weighted Relative Mean Deviation $D(\alpha) = \left(\sum_{i=1}^n \sigma_i \left|y_i(\alpha) - \mu_{y(\alpha)}\right|\right)/\mu_{y(\alpha)}$. Taking $\alpha = 0$, we obtain $\mu_y = \sum_{i=1}^n \sigma_i y_i$ and $D = \left(\sum_{i=1}^n \sigma_i \left|y_i - \mu_y\right|\right)/\mu_y$. Using these definitions, the weight of individual i is equal to:

$$w_i^L(\alpha) = n\sigma_i . \frac{y_i(\alpha) - \mu_{y(\alpha)}}{\mu_{y(\alpha)}} . \frac{1}{\phi(\alpha)} \quad \text{(A.4)}$$

which for the basic version ($\alpha = 0$) reduces to:

$$w_i^L = n\sigma_i . \frac{y_i - \mu_y}{\mu_y} \quad \text{(A.5)}$$

References

[1] Abul Naga, R.H. and P.-Y. Geoffard (2006), "Decomposition of bivariate inequality indices by attributes", *Economics Letters*, 90: 362-367.

[2] Australian Bureau of Statistics (2013), *Household Income and Income Distribution, Australia, 2011-12*, Canberra, ABS, Catalogue No. 6523.0.

[3] Bhattacharya, N. and B. Mahalanobis (1967), "Regional disparities in household consumption in India", *Journal of the American Statistical Association*, 62: 143-161.

[4] Blinder, A. (1973), "Wage discrimination: reduced form and structural estimates", *Journal of Human Resources*, 8: 945-957.

[5] Bourguignon, F. (1979), "Decomposable income inequality measures", *Econometrica*, 47: 901-920.

[6] Clarke, P., U.-G. Gerdtham and L.B. Connelly (2003), "A note on the decomposition of the health concentration index", *Health Economics*, 12: 511-516.

[7] Donaldson, D. and J.A. Weymark (1980), "A single-parameter generalization of the Gini indices of inequality", *Journal of Economic Theory*, 22: 67-86.

[8] Erreygers, G. (2009), "Correcting the concentration index", *Journal of Health Economics*, 28: 504-515.

[9] Erreygers, G. (2013), "A dual Atkinson measure of socioeconomic inequality of health", *Health Economics*, 22: 466-479.

[10] Erreygers, G., P. Clarke and T. Van Ourti (2012), "'Mirror, mirror, on the wall, Who in this land is fairest of all?' – Distributional sensitivity in the measurement of socioeconomic inequality of health", *Journal of Health Economics*, 31: 257-270.

[11] Erreygers, G. and R. Kessels (2013), "Regression-based decompositions of rank-dependent indicators of socioeconomic inequality of health", in: P. Rosa Dias and O. O'Donnell (Eds.), *Health and Inequality* (Research on Economic Inequality, Volume 21), Emerald, pp. 227-259.

[12] Erreygers, G. and R. Kessels (2015), *Socioeconomic Status and Health: A New Approach to the Measurement of Bivariate Inequality*, Antwerp, University of Antwerp, Faculty of Applied Economics, Working Paper 2015/017.

[13] Erreygers, G. and T. Van Ourti (2011), "Measuring socioeconomic inequality in health, health care, and health financing by means of rank-dependent indices: A recipe for good practice", *Journal of Health Economics*, 30: 685-694.

[14] Kakwani, N.C. (1980), *Income Inequality and Poverty. Methods of Estimation and Policy Applications*, Oxford, Oxford University Press.

[15] Kessels, R. and G. Erreygers (2015), *A Unified Structural Equation Modelling Approach for the Decomposition of Rank-Dependent Indicators of Socioeconomic Inequality of Health*, Helsinki, United Nations University, UNU-WIDER, WIDER Working Paper 2015/017.

[16] Lambert, P.J. and J.R. Aronson (1993), "Inequality decomposition analysis and the Gini coefficient revisited", *Economic Journal*, 103: 1221-1227.

[17] Lambert, P.J. and G. Lanza (2006), "The effect on inequality of changing one or two incomes", *Journal of Economic Inequality*, 4: 253-277.

[18] Makdissi, P., D. Sylla and M. Yazbeck (2013), "Decomposing health achievement and socioeconomic health inequalities in presence of multiple categorical information", *Economic Modelling*, 35: 964-968.

[19] Mehran, F. (1976), "Linear measures of income inequality", *Econometrica*, 44: 805-809.

[20] Mookherjee and Shorrocks (1982), "A decomposition analysis of the trend in UK income inequality", *Economic Journal*, 92: 886-902.

[21] Oaxaca, R.L. (1973), "Male-female wage differentials in urban labor markets", *International Economic Review*, 14: 693-709.

[22] O'Donnell, O., E. Van Doorslaer and T. Van Ourti (2015), "Health and inequality", in: A.B. Atkinson and F.J. Bourguignon (Eds.), *Handbook of Income Distribution*, Amsterdam, Elsevier, Volume 2, Part B, chapter 18, pp. 1419-1533.

[23] O'Donnell, O., E. Van Doorslaer and A. Wagstaff (2012), "Decomposition of inequalities in health and health care", in: A.M. Jones (Ed.), *The Elgar Companion to Health Economics, Second Edition*, Cheltenham, Edward Elgar, chapter 17, pp. 179-191.

[24] Pereira, J.A. (1998), "Inequality in infant mortality in Portugal, 1971-1991", in: Zweifel, P. (Ed.), *Health, the Medical Profession, and Regulation* (Developments in Health Economics and Public Policy, vol. 6). Kluwer, Boston/Dordrecht/London, pp. 75-93.

[25] Rao, V. M. (1969), "Two decompositions of concentration ratio", *Journal of the Royal Statistical Society*, Series A (General), 132: 418-425.

[26] Shorrocks, A.F. (1980), "The class of additively decomposable inequality measures", *Econometrica*, 48: 613-625.

[27] Van Doorslaer, E. and T. Van Ourti (2011), "Measuring inequality and inequity in health and health care", in: S. Glied and P.C. Smith (Eds.), *The Oxford Handbook of Health Economics*, Oxford, Oxford University Press, chapter 35, pp. 837-869.

[28] Van Ourti T., G. Erreygers and P. Clarke (2014), "Measuring equality and equity in health and health care", in: A.J. Culyer (Ed.), *Encyclopedia of Health Economics*, San Diego: Elsevier, Vol. 2, pp. 234-239.

[29] Wagstaff, A. (2002), "Inequality aversion, health inequalities, and health achievement", *Journal of Health Economics*, 21: 627-641.

[30] Wagstaff, A. (2005a), "The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality", *Health Economics*, 14: 429-432.

[31] Wagstaff, A. (2005b), "Inequality decomposition and geographic targeting with applications to China and Vietnam", *Health Economics*, 14: 649-635.

[32] Wagstaff, A., P. Paci and E. Van Doorslaer (1991), "On the measurement of inequalities in health", *Social Science and Medicine*, 33: 545–57.

[33] Wagstaff, A., E. Van Doorslaer and N. Watanabe (2003), "On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam", *Journal of Econometrics*, 112: 207-223.

[34] Yitzhaki, S. (1983), "On an extension of the Gini inequality index", *International Economic Review*, 24: 617-628.