DEPARTMENT OF ECONOMICS

Regression-Based Decompositions of Rank-Dependent Indicators of Socioeconomic Inequality of Health

Guido Erreygers & Roselinde Kessels

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

Stadscampus
Prinsstraat 13, B.226
BE-2000 Antwerpen
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
www.ua.ac.be/tew

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF ECONOMICS

Regression-Based Decompositions of Rank-Dependent Indicators of Socioeconomic Inequality of Health

Guido Erreygers & Roselinde Kessels

RESEARCH PAPER 2013-007
APRIL 2013

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@ua.ac.be

The papers can be also found at our website:
www.ua.ac.be/tew (research > working papers) &
www.repec.org/ (Research papers in economics - REPEC)

D/2013/1169/007

Regression-Based Decompositions of Rank-Dependent Indicators of Socioeconomic Inequality of Health

Guido Erreygers and Roselinde Kessels

April 22, 2013

Abstract

In this paper we explore different ways to obtain decompositions of rank-dependent indices of socioeconomic inequality of health, such as the Concentration Index. Our focus is on the regression-based type of decomposition. Depending on whether the regression explains the health variable, or the socioeconomic variable, or both, a different decomposition formula is generated. We illustrate the differences using data from the Ethiopia 2011 Demographic and Health Survey (DHS).

Acknowledgements: We thank Tom Van Ourti and Philip Clarke for discussions related to decomposition analysis, and Ellen Van de Poel for assistance with regard to the Ethiopian data. We are also grateful to seminar participants at the Centre of Health Economics Research (COHERE) of the University of Southern Denmark, and the Department of Economics of the University of Antwerp.

Keywords: Inequality measurement, Concentration index, Decomposition methods.

JEL Classification Number: D63, I00

Address of the authors: Department of Economics, University of Antwerp, City Campus, Prinsstraat 13, 2000 Antwerpen, Belgium, guido.erreygers@ua.ac.be; Department of Engineering Management, University of Antwerp, City Campus, Prinsstraat 13, 2000 Antwerpen, Belgium, roselinde.kessels@ua.ac.be

1 Introduction

Inequality decomposition techniques have been developed and used extensively over the last four to five decades. In the income inequality literature various methods have been proposed to decompose inequality by subgroups as well as by income sources (see, e.g., Rao, 1969 and Bourguignon, 1979 for early contributions). Typical for subgroup decompositions is that they involve partitions of the population into distinct subsets, such as urban and rural population, or male and female population. In the literature on subgroup decomposability the focus has been on additively decomposable indicators, i.e. inequality measures which can be written as a weighted sum of within-group inequality levels and a between-group term (see, e.g., Shorrocks, 1980), but some attention has also been paid to inequality measures satisfying more general decomposability criteria (see, e.g., Shorrocks, 1984). Income source decompositions refer to divisions of income into different components, such as wage and non-wage income. For these decompositions, the main issue has been whether it is possible to identify the proportional contribution of each income component to the measured level of inequality (see, e.g., Shorrocks, 1982).

Both decompositions are based on pre-determined identities which split either the population or individual incomes into distinct categories. By contrast, regression-based decompositions rely on estimations of incomes. The identities are replaced by regressions of which the independent variables are the determinants of income, such as education, sex, and age. Because of the presence of an error term in the regression equation, these decompositions typically end up with an unexplained residual. The origins of this approach can be traced back to the influential work of Oaxaca (1973), but especially Morduch and Sicular (2002) and Fields (2003) have developed the technique further by building on insights from the literature on income source decompositions. Cowell and Fiorio (2011) provide a good overview of recent contributions in this field.

Some of the income inequality decomposition methods have also been applied to measures of socioeconomic inequality of health. Simple decomposition formulas by both subgroups and health components were suggested by Clarke, Gerdtham and Connelly (2003). Regression-based decomposition methods were introduced by Gravelle (2003) and Wagstaff, Van Doorslaer and Watanabe (2003). Using results on the decomposition of the Gini coefficient obtained by Podder (1993), Wagstaff, Van Doorslaer and Watanabe (2003) not only proposed a procedure to decompose the Concentration Index into contributing factors, but also indicated how the decomposition could serve as a framework to analyse changes in socioeconomic inequality

over time. Their results have been used and extended in many ways (see, e.g., Jones and López Nicolás, 2004; O'Donnell, Van Doorslaer and Wagstaff, 2006; O'Donnell et al., 2008, chapter 13; Van Ourti, Van Doorslaer and Koolman, 2009) and led to numerous empirical applications (see, e.g., Hosseinpoor et al., 2006; Van de Poel et al., 2007). Some of this work is covered in a recent survey by Van Doorslaer and Van Ourti (2011).

In this paper we focus on regression-based decompositions of rank-dependent indicators of socioeconomic inequality of health, of which the Concentration Index is the best known example. Our starting point is that inequality of income is of a different nature than socioeconomic inequality of health: while the former is univariate, the latter is bivariate. In other words, whereas income inequality indicators measure the degree of inequality of incomes as such, socioeconomic inequality indicators measure the degree of *correlation* between socioeconomic status and health. This difference in character is visible in the range of values of the indicators: univariate indicators are always nonnegative (e.g., limited to the interval $[0, 1]$, with 0 indicating minimum inequality and 1 maximum inequality), but bivariate indicators can be both negative and positive (e.g., limited to the interval $[-1, +1]$, with -1 indicating an extreme pro-poor distribution of health and $+1$ an extreme pro-rich distribution of health). We fear that the difference is not always properly acknowledged, although there are exceptions (see, e.g., Abul Naga and Geoffard, 2006). The challenge, therefore, is whether it is possible to find decompositions which explain the degree of correlation between two variables rather than the degree of inequality in one variable.

After a preliminary section, we first deal with one-dimensional decompositions, including the Wagstaff-Van Doorslaer-Watanabe decomposition (section 3). Then we explore various two-dimensional decompositions (section 4), and make a comparison of both (section 5). By means of an empirical analysis of stunting in Ethiopia we illustrate the differences between the main decompositions (section 6). We discuss the results (section 7) and draw a number of conclusions (section 8).

2 Preliminaries

Consider a population consisting of n individuals. The health level of individual i is represented by the real number h_i. The health variable is assumed to be either a ratio-scale variable which takes nonnegative values only, or a cardinal variable with a finite lower bound. The average health level in the population is equal to $\mu_h = \frac{1}{n}\sum_{i=1}^{n} h_i$.

As argued by Erreygers and Van Ourti (2011), the use of the health Con-

centration Index can be defended when we are dealing with a ratio-scale health variable which is unbounded, i.e. which does not have a finite upper bound. However, when we are dealing with a variable which has a finite upper bound, a modified version is called for. For this situation, Wagstaff (2005) and Erreygers (2009) each proposed a variant of the Generalized Concentration Index.

All of these indices are rank-dependent indices in the sense that they are in essence weighted sums of health levels with the weights determined by socioeconomic ranks. The socioeconomic rank of individual i is determined by her position according to the variable which is supposed to measure socioeconomic well-being, e.g. income. Let the value of this variable for individual i be y_i. Then the natural number $r_i(y)$ measures the position of individual i in the rank-order according to variable y, with the rank $r_i(y) = 1$ assigned to the person who is least well-off, and the rank $r_i(y) = n$ assigned to the person who is most well-off. (In what follows we will simplify $r_i(y)$ to r_i if it is obvious that y is the variable used for ranking.) In the case of ties, we assign to every individual of the tied group the average rank of the group. Over the population as a whole the average rank is $\mu_r = \frac{n+1}{2}$. The fractional rank f_i is defined as $f_i \equiv \frac{1}{n}\left(r_i - \frac{1}{2}\right)$, and varies between $\frac{1}{2n}$ and $1 - \frac{1}{2n}$. The average fractional rank is $\mu_f = \frac{1}{2}$. Finally, the deviation of the fractional rank of individual i from the average fractional rank, denoted as $d_i \equiv f_i - \mu_f$, has an average of $\mu_d = 0$.

The Generalized health Concentration Index $GC(h; y)$, or more simply GC whenever it is clear that the health variable is h and the socioeconomic variable y, is defined as:

$$GC = \frac{2}{n} \sum_{i=1}^{n} h_i d_i \tag{1}$$

The standard health Concentration Index, $C(h; y) = C$, as well as the indices introduced by Wagstaff (2005), $W(h; y) = W$, and by Erreygers (2009), $E(h; y) = E$, can be expressed as simple functions of GC:

$$C = \frac{1}{\mu_h} GC \tag{2}$$

$$W = \frac{b_h - a_h}{(b_h - \mu_h)(\mu_h - a_h)} GC \tag{3}$$

$$E = \frac{4}{b_h - a_h} GC \tag{4}$$

where a_h and b_h stand for the lower and upper bounds of the health variable. In the remainder of the paper we will concentrate on the decomposition of

the Generalized Concentration Index; formulas (2), (3) and (4) can be used to convert the outcome into corresponding decompositions of the other rank-dependent indices.

To end this preliminary section, it may be useful to point out a well-known connection between the rank-dependent indices and the covariance concept. Since $Cov(h,d) = \frac{1}{n}\sum_{i=1}^{n} h_i d_i - \mu_h \mu_d$ and $\mu_d = 0$, it follows that we have:

$$GC = 2Cov(h,d) \tag{5}$$

Both (1) and (5) can be used to generate decompositions of the Generalized Concentration Index. Although one expects to obtain the same decompositions irrespective of which of the two formulas is used, we will show below that this need not be the case.

3 One-dimensional Decompositions

We begin our analysis with the case in which only one of the two variables is subject to a regression. This has been the dominant approach in the literature on regression-based decompositions of the Concentration Index. Following Wagstaff, Van Doorslaer and Watanabe (2003) health has been unanimously preferred as the relevant variable for regression. We therefore first recapitulate the results of this health-oriented decomposition and then introduce rank-oriented decompositions.

3.1 The Health-oriented Decomposition

Wagstaff, Van Doorslaer and Watanabe (2003), who have pioneered this type of decomposition analysis, start from the assumption that the health variable h is determined by a number of explanatory variables $x_1, x_2, ..., x_k$. Suppose that we have the following linear regression equation:

$$h_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \varepsilon_i \tag{6}$$

where ε_i is an error term. Substituting the right-hand side of (6) for h_i in (1) we obtain:

$$GC = \frac{2}{n}\sum_{i=1}^{n} \left[\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \varepsilon_i\right] d_i \tag{7}$$

$$= \beta_0 \frac{2}{n}\sum_{i=1}^{n} d_i + \sum_{j=1}^{k} \beta_j \left(\frac{2}{n}\sum_{i=1}^{n} x_{j,i} d_i\right) + \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i d_i \tag{8}$$

Since $\sum_{i=1}^{n} d_i = 0$, the first term vanishes. Given that $\frac{1}{n}\sum_{i=1}^{n} x_{j,i}d_i = Cov(x_j, d)$ and $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i d_i = Cov(\varepsilon, d)$, this leads to the following decomposition:

$$GC = 2\sum_{j=1}^{k} \beta_j Cov(x_j, d) + 2Cov(\varepsilon, d) \tag{9}$$

This decomposition splits the Generalized Concentration Index into two components: an explained part and a residual term. The explained component consists of a sum of k contributions, one for each explanatory variable. If the x_j variables are ratio-scale or cardinal variables, we can apply definition (1) to each of the k individual terms. In fact, in that case the contribution of the explanatory variable x_j to socioeconomic inequality, $2\beta_j Cov(x_j, d)$, can also be expressed as a Generalized Concentration Index, since it is equal to $\beta_j GC(x_j; y)$. If the error term of (6) were uncorrelated with the socioeconomic rank, one would expect the residual component $Cov(\varepsilon, d)$ to be close to zero. In applied work, however, this is often not the case.

3.2 Rank-oriented Decompositions

Instead of starting from an equation which predicts health levels, we could also start from an equation which predicts the socioeconomic ranks. Two methods suggest themselves. The first consists of predicting the levels of the variable which defines the socioeconomic ranks. Assuming that y is determined by the variables $z_1, z_2, ..., z_q$, the first step would be to estimate the linear equation:

$$y_i = \gamma_0 + \gamma_1 z_{1,i} + \gamma_2 z_{2,i} + ... + \gamma_q z_{q,i} + \eta_i \tag{10}$$

where η_i is an error term. The predicted value of y_i, i.e. $\widehat{y}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 z_{1,i} + \widehat{\gamma}_2 z_{2,i} + ... + \widehat{\gamma}_q z_{q,i}$, could then be used to generate the predicted rank of person i, $\widehat{r}_i = r_i(\widehat{y})$. Let us call the difference between the actual and the predicted ranks the unexplained rank deviation, and let us denote it as ρ_i. Given that $r_i = \widehat{r}_i + \rho_i$, we also have, using an obvious notation, $f_i = \widehat{f}_i + \omega_i$ and $d_i = \widehat{d}_i + \omega_i$, where $\omega_i = \rho_i/n$. Substituting this into (1) we obtain:

$$GC = \frac{2}{n}\sum_{i=1}^{n} h_i \left(\widehat{d}_i + \omega_i\right) \tag{11}$$

and so the Generalized Concentration Index can be decomposed into two components, one explained and the other unexplained:

$$GC = 2Cov(h, \widehat{d}) + 2Cov(h, \omega) \tag{12}$$

The drawback of this first method, which determines the socioeconomic ranks indirectly, is that it allows only a very rough decomposition of the index. As an alternative, we could try to determine the socioeconomic ranks directly. Hence, assuming that the variables $z_1, z_2, ..., z_q$ are the relevant variables, instead of estimating equation (10) we could start from an equation which estimates the fractional rank deviations:

$$d_i = \alpha_0 + \alpha_1 z_{1,i} + \alpha_2 z_{2,i} + ... + \alpha_q z_{q,i} + \xi_i \tag{13}$$

where ξ_i is an error term. Substituting this into (1) we now obtain the following result:

$$GC = 2\alpha_0 \mu_h + 2 \sum_{g=1}^{q} \alpha_g \sum_{i=1}^{n} h_i z_{g,i} + 2 \sum_{i=1}^{n} h_i \xi_i \tag{14}$$

This equation suggests a three-component decomposition:

$$GC = 2\alpha_0 \mu_h + 2 \sum_{g=1}^{q} \alpha_g \left[Cov(h, z_g) + \mu_h \mu_{z_g} \right] + 2Cov(h, \xi) \tag{15}$$

We now have: (1) a constant term, $2\alpha_0 \mu_h$; (2) a sum of contributions of the q explanatory variables, with the contribution of variable g equal to $2\alpha_g \left[Cov(h, z_g) + \mu_h \mu_{z_g} \right]$; and (3) an unexplained component, $2Cov(h, \xi)$.

One may feel uncomfortable with this decomposition because of the presence of a constant term, which has no obvious interpretation. Following (5) rather than (1) suggests a way out. Since $Cov(h, d) = Cov(h, \alpha_0 u + \alpha_1 z_1 + \alpha_2 z_2 + ... + \alpha_q z_q + \xi)$, where u is a vector of ones, it follows that $Cov(h, d) = \alpha_0 Cov(h, u) + \alpha_1 Cov(h, z_1) + \alpha_2 Cov(h, z_2) + ... + \alpha_q Cov(h, z_q) + Cov(h, \xi)$. Knowing that $Cov(h, u) = 0$, we arrive at the following decomposition:

$$GC = 2 \sum_{g=1}^{q} \alpha_g Cov(h, z_g) + 2Cov(h, \xi) \tag{16}$$

The modified decomposition is written as a sum of two components, one explained and the other unexplained. The contribution of variable z_g to the explained component is now equal to $2\alpha_g Cov(h, z_g)$. Actually, we could have derived (16) from (15); we have in fact $\alpha_0 + \sum_{g=1}^{q} \alpha_g \mu_{z_g} = 0$, which follows from (13) if one takes into account that $\mu_d = 0$ and $\mu_\xi = 0$. Comparing (16) to (15), it is as if the value of the constant term is distributed over the contributions of the explanatory variables, with $-2\alpha_g \mu_h \mu_{z_g}$ being added to the contribution of variable z_g.

A special case occurs when all the independent variables $z_1, z_2, ..., z_q$ are themselves rank variables, or more specifically fractional rank deviation variables. Since all dependent and independent variables of (13) then have zero means, it follows that $\alpha_0 = -\mu_\xi$. Hence, if we use a regression technique for which $\mu_\xi = 0$, we will find that $\alpha_0 = 0$. It follows that in this case the two decompositions (15) and (16) will coincide. Since every z_j is a rank variable, each explained component $2\alpha_g Cov(h, z_g)$ can be interpreted as a Generalized Concentration Index, but with z_j rather than d used to calculate the index.

Decomposition (16) has a similar structure as the Wagstaff-Van Doorslaer-Watanabe decomposition (9). In each case we arrive at an expression which decomposes the Generalized Concentration Index into a sum of q explained components, with each of these components equal to a covariance weighted by a regression coefficient, and a residual or unexplained component, which is also a covariance. Although decomposition (15) is slightly different, it has a similar residual term. If the error term ξ were uncorrelated with the health variable h, the covariance $Cov(h, \xi)$ would be close to zero and the residual term very small.

4 From One-dimensional to Two-dimensional Decompositions

So far we have limited ourselves to one-dimensional decompositions. These are based on regressions of only one of the two variables under consideration. It should be remembered, however, that we are interested in the *correlation* between the two variables. It may be doubted whether focusing on one dimension only is enough to fully understand how the two dimensions are related. Therefore, it seems appropriate to extend the analysis and to adopt a framework which allows one to look at both variables simultaneously.

Three approaches suggest themselves, and we will explore each one of them briefly. The first approach consists of making a super-decomposition based on two separate regressions, one for each variable. In the second approach the two regressions are made simultaneously, using a common set of independent variables. The final approach looks for a direct way to explain the correlation between the two variables.

4.1 A Combined Super-decomposition

Let us assume that each of the two variables of interest can be explained by a separate regression equation. More specifically, let health levels be

predicted by (6) and socioeconomic ranks by (13). A combination of these two equations results in what we will call a combined super-decomposition.

Once again, two paths can be followed to obtain a decomposition formula. If we start from (1) the Generalized health Concentration Index will be expressed as:

$$GC = \frac{2}{n}\sum_{i=1}^{n}\left[\beta_0 + \sum_{j=1}^{k}\beta_j x_{j,i} + \varepsilon_i\right]\left[\alpha_0 + \sum_{g=1}^{q}\alpha_g z_{g,i} + \xi_i\right] \tag{17}$$

Working out the terms and assuming that $\mu_\varepsilon = \mu_\xi = 0$, we can expand this to:

$$\begin{aligned} GC &= 2\beta_0\alpha_0 + 2\alpha_0\sum_{j=1}^{k}\beta_j\mu_{x_j} + 2\beta_0\sum_{g=1}^{q}\alpha_g\mu_{z_g} \\ &\quad +2\sum_{j=1}^{k}\sum_{g=1}^{q}\beta_j\alpha_g Cov(x_j, z_g) \\ &\quad +2\left[\sum_{j=1}^{k}\beta_j Cov(x_j,\xi) + \sum_{g=1}^{q}\alpha_g Cov\left(\varepsilon, z_g\right) + Cov(\varepsilon,\xi)\right] \end{aligned} \tag{18}$$

In comparison to what we obtained before, this result looks awfully complicated. Adopting a two-dimensional approach leads to a dramatic increase in the number of terms. We can simplify the decomposition a bit by making the plausible assumption that the $k+q$ covariance terms $Cov(x_j,\xi)$ and $Cov\left(\varepsilon, z_g\right)$ are all equal to zero. This gives:

$$\begin{aligned} GC &= 2\beta_0\alpha_0 + 2\alpha_0\sum_{j=1}^{k}\beta_j\mu_{x_j} + 2\beta_0\sum_{g=1}^{q}\alpha_g\mu_{z_g} \\ &\quad +2\sum_{j=1}^{k}\sum_{g=1}^{q}\beta_j\alpha_g Cov(x_j, z_g) + 2Cov(\varepsilon,\xi) \end{aligned} \tag{19}$$

which is only marginally less complicated.

As before, starting from (5) generates a more simple decomposition. Since $Cov(h,d) = Cov(\beta_0 u + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon, \alpha_0 u + \alpha_1 z_1 + \alpha_2 z_2 + ... + \alpha_q z_q + \xi)$, we obtain:

$$GC = 2\sum_{j=1}^{k}\sum_{g=1}^{q}\beta_j\alpha_g Cov(x_j, z_g) + 2Cov(\varepsilon,\xi) \tag{20}$$

Instead of just k or q explained contributions, we now have kq terms which may be said to capture the explained component of the Generalized Concentration Index. It is, however, much harder than before to give a clear interpretation to the many terms involved in the decomposition. For instance, saying that the total explained effect of variable x_j is equal to $2\beta_j \sum_{g=1}^{q} \alpha_g Cov(x_j, z_g)$ and that of variable z_g is equal to $2\alpha_g \sum_{j=1}^{k} \beta_j Cov(x_j, z_g)$ would lead to double counting.

4.2 A Simultaneous Super-decomposition

A special case would occur if the set of variables which determine the health levels – the x_j's – coincided with the set of variables which determine the socioeconomic positions – the z_g's. Suppose we have a common set of p variables $s_1, s_2, \ldots, s_p$ which determine both the health levels h and the fractional rank deviations d. Then we could run a bivariate regression of the following form:

$$h_i = \lambda_0 + \lambda_1 s_{1,i} + \lambda_2 s_{2,i} + ... + \lambda_p s_{p,i} + \psi_i \quad (21)$$

$$d_i = \pi_0 + \pi_1 s_{1,i} + \pi_2 s_{2,i} + ... + \pi_p s_{p,i} + \chi_i \quad (22)$$

where ψ_i and χ_i are error terms. We take it for granted that $\mu_\psi = \mu_\chi = 0$, and moreover that the $2p$ covariances $Cov(s_j, \chi)$ and $Cov(\psi, s_j)$ are zero.

We can again start from either (1) or (5). In the first case, the application of formula (19) gives the following expression:

$$\begin{aligned} GC = {} & 2\lambda_0\pi_0 + 2\sum_{j=1}^{p}(\lambda_0\pi_j + \lambda_j\pi_0)\mu_{s_j} \\ & + 2\sum_{j=1}^{p}\sum_{g=1}^{p}\lambda_j\pi_g\left[Cov(s_j, s_g) + \mu_{s_j}\mu_{s_g}\right] + 2Cov(\psi, \chi) \end{aligned} \quad (23)$$

Although this might already be considered as a decomposition formula, we propose to group the terms slightly differently. The second component of the right-hand side of (23) consists of a sum of p terms, one for each independent variable s_j, where $j = 1, ..., p$. The third component, by contrast, has p^2 terms, one for each pair of independent variables (s_j, s_g), where $j, g = 1, ..., p$. Of these p^2 terms, exactly p refer to only one variable. Let us add these terms to each of the corresponding terms of the second component. Given that

$Cov(s_j, s_j) = Var(s_j)$ and $Cov(s_j, s_g) = Cov(s_g, s_j)$, this gives the following:

$$\begin{aligned} GC &= 2\lambda_0\pi_0 + 2\sum_{j=1}^{p}\left\{(\lambda_0\pi_j + \lambda_j\pi_0)\mu_{s_j} + \lambda_j\pi_j\left[Var(s_j) + (\mu_{s_j})^2\right]\right\} \\ &\quad +2\sum_{j=1}^{p}\sum_{g=j+1}^{p}\left\{(\lambda_j\pi_g + \lambda_g\pi_j)\left[Cov(s_j, s_g) + \mu_{s_j}\mu_{s_g}\right]\right\} \\ &\quad +2Cov(\psi, \chi) \end{aligned} \tag{24}$$

In other words, we now have: (1) a constant term, $2\lambda_0\pi_0$; (2) p single-variable terms $2(\lambda_0\pi_j + \lambda_j\pi_0)\mu_{s_j} + \lambda_j\pi_j\left[Var(s_j) + (\mu_{s_j})^2\right]$ which might be said to capture the direct effect of the p independent variables; (3) $\frac{p(p-1)}{2}$ two-variable terms $2(\lambda_j\pi_g + \lambda_g\pi_j)\left[Cov(s_j, s_g) + \mu_{s_j}\mu_{s_g}\right]$ which capture the correlation structure between the independent variables; and (4) an unexplained component which is equal to the covariance between the two error terms. For a well-specified model one expects the correlation effects to be small relative to the direct effects.

If we start from (5) rather than (1), we have to apply formula (20). Isolating the terms of the first component which involve only one of the independent variables, and keeping in mind that all $Cov(s_j, \chi)$ and $Cov(\psi, s_j)$ are by assumption equal to zero, we arrive at the following decomposition:

$$\begin{aligned} GC &= 2\sum_{j=1}^{p}\lambda_j\pi_j Var(s_j) + 2\sum_{j=1}^{p}\sum_{g=j+1}^{p}(\lambda_j\pi_g + \lambda_g\pi_j)Cov(s_j, s_g) \\ &\quad +2Cov(\psi, \chi) \end{aligned} \tag{25}$$

In comparison to decomposition (24) it is as if the value of the constant term has been distributed over both the direct and combined or correlated contributions.

4.3 A Direct Explanation of the Correlation Between Health and Socioeconomic Status

The two super-decompositions outlined above are based on two regressions, one for each of the two dimensions involved. Each regression equation explains the variability in one dimension only. The product of the two regression equations is then supposed to provide an explanation of the correlation between the two dimensions. But does this indirect procedure really work? Should we not aim for a direct explanation of the observed correlation between the two dimensions?

The question is whether it is possible to do some kind of regression where the dependent variable measures the variability of the *correlation* between the two variables, rather than the variability in one of them only. There is no natural unit to measure correlation at the individual level. But knowing that $GC = \frac{2}{n}\sum_{i=1}^{n} h_i d_i$, a proxy variable suggests itself. Let us define the composite variable $v_i \equiv h_i d_i$ and treat it as a measure of correlation between h_i and d_i. Assuming that the independent variables $s_1, s_2, \ldots, s_p$ are the determining factors of the degree of correlation, we can do the following linear regression to explain this new composite variable:

$$v_i = \phi_0 + \phi_1 s_{1,i} + \phi_2 s_{2,i} + ... + \phi_p s_{p,i} + \zeta_i \tag{26}$$

where ζ_i is an error term. Given that $GC = \frac{2}{n}\sum_{i=1}^{n} v_i$ and assuming that $\mu_\varsigma = 0$, it follows that we have:

$$GC = 2\phi_0 + 2\phi_1 \mu_{s_1} + 2\phi_2 \mu_{s_2} + ... + 2\phi_p \mu_{s_p} \tag{27}$$

We therefore arrive at a decomposition consisting of a constant term $2\phi_0$, which again is difficult to interpret, and p terms $2\phi_j \mu_{s_j}$, of which the jth term may be interpreted as the contribution of factor s_j to GC. In contrast to the previous decompositions, there is no unexplained component.

5 A Comparison

Now that we have identified several decomposition possibilities, let us take a step back and see how they compare to one another. We no longer take into consideration the very rough decomposition (12) based on the indirectly estimated ranks derived from the regression of the underlying socioeconomic variable. As a matter of fact, in order to make the comparison meaningful we limit ourselves to decompositions based on regressions with a common set of independent variables $s_1, s_2, ..., s_p$. We therefore assume that $\{x_1, x_2, ..., x_k\} = \{z_1, z_2, ..., z_q\} = \{s_1, s_2, ..., s_p\}$. It follows that the coefficients and the error terms of the original health regression (6) and of the original fractional rank deviation regression (13) coincide with those of the bivariate regression equations (21)–(22), i.e. $\beta = \lambda$, $\alpha = \pi$, $\varepsilon = \psi$ and $\xi = \chi$.

There are six remaining decompositions to be compared. Three of these are one-dimensional:

(I) the health-oriented decomposition based on regression (6) and represented by (9);

(IIa) the rank-oriented decomposition based on regression (13) and represented by (15), which has a constant term;

(IIb) the rank-oriented decomposition based on regression (13) and represented by (16), without a constant term.
The other three are two-dimensional:

(IIIa) the super-decomposition based on the bivariate regression model (21)–(22) and represented by (24), which has a constant term;

(IIIb) the super-decomposition based on the bivariate regression model (21)–(22) and represented by (25), without a constant term;

(IV) the decomposition based on the regression of the health-fractional rank deviation product (26) and represented by (27).

Table 1 summarizes the components of each decomposition.

*** Insert Table 1 around here ***

Some observations are in order with regard to these decompositions. First, it can be shown (see Appendix) that the residual terms of decompositions (I), (IIa), (IIb), (IIIa) and (IIIb) are all the same. This remarkable result comes from the fact that the two regressions on which they are based have exactly the same set of independent variables. Second, the p individual contributions of decompositions (I) and (IIb) can be related to the direct and combined contributions of decomposition (IIIb). In fact, our assumption that the $2p$ covariance terms $Cov(s_j, \chi)$ and $Cov(\psi, s_j)$ are zero implies:

$$2\lambda_j Cov(s_j, d) = 2\lambda_j \sum_{g=1}^{p} \pi_g Cov(s_j, s_g) \tag{28}$$

$$= 2\lambda_j \pi_j Var(s_j) + 2 \sum_{g=1, g\neq j}^{p} \lambda_j \pi_g Cov(s_j, s_g) \tag{29}$$

$$2\pi_j Cov(h, s_j) = 2\pi_j \sum_{g=1}^{p} \lambda_g Cov(s_g, s_j) \tag{30}$$

$$= 2\lambda_j \pi_j Var(s_j) + 2 \sum_{g=1, g\neq j}^{p} \lambda_g \pi_j Cov(s_j, s_g) \tag{31}$$

A slightly more complex relationship holds between the terms of decompositions (I), (IIa) and (IIIa).

6 An Empirical Illustration

6.1 Data Description

The data come from the 2011 Demographic and Health Survey (DHS) of Ethiopia and are confined to children under the age of five.

First, we constructed the response variables, the health variable h, the fractional rank deviation d and the composite variable v, for use in the six decompositions. The health variable h we have chosen is actually an ill-health variable: the degree of stunting or malnutrition. It is defined on the unit interval, i.e. $a_h = 0$ and $b_h = 1$, and provides information on the depth of malnutrition with children. We obtained this measure from the child's height-for-age standard deviation or z-score which is the difference between the height of a child and the median height of a child of the same age and sex in a well-nourished reference population divided by the standard deviation in the reference population. We used the new WHO child growth population as reference population. We measured the degree of stunting relative to the threshold of minus two standard deviations of the median of the reference population. Children with a z-score greater than this threshold are not stunted and are assigned a zero degree value. The other children are stunted and are assigned a value in the unit interval that is proportional to the magnitude of their z-score with a z-score of minus six standard deviations corresponding to the maximum value of one. In total, taking into account the sample weights provided by DHS, we found that 44% of the children in our dataset (see below for further description) are stunted. To obtain the fractional rank deviation d, we ranked the children's households according to their wealth status. We used the wealth indices constructed by DHS from a principal component analysis on all household living conditions and assets. We computed the fractional rank deviation taking into account the sample weights so that, in effect, the variable d stands for the weighted fractional rank deviation. To create the composite variable v, we multiplied the variables h and d for all children in our dataset.

Second, we selected a set of nine explanatory variables for use in the decompositions. Based on previous stunting regressions performed by Wagstaff, Van Doorslaer and Watanabe (2003) and Van de Poel et al. (2007), we included a number of child-level characteristics such as age and sex of the child, maternal characteristics such as education of the mother and her partner or husband and household-level characteristics such as urban or rural residence, time to a water source, access to safe drinking water and satisfactory sanitation. In addition to that, we specified the child's age nonlinearly in the regression models using a squared term. Because Wagstaff, Van Doorslaer

and Watanabe (2003) found a significant inverted u-shaped relationship for child's age in their stunting regressions for Vietnam, we expected to obtain a similar result. We mean-centered the squared term, however, to remove multicollinearity with the linear term. Furthermore, we defined the safe drinking water and satisfactory sanitation variables along the lines proposed by the WHO and UNICEF. We identified the following sources of water supply as safe drinking water: piped water (piped into dwelling, piped into yard or plot, or public tap), water from a protected well, tube well or borehole, water from a protected spring and rainwater. We identified the following sanitation infrastructure as satisfactory sanitation: a flush toilet (flush to piped sewer system, septic tank or pit latrine), a pit latrine with slab, a Ventilated Improved Pit (VIP) latrine and a composting toilet. Note that, as opposed to the regressions made by Wagstaff, Van Doorslaer and Watanabe (2003) and Van de Poel et al. (2007), we did not include a wealth-related variable in our set of explanatory variables. The reason is that we used the wealth indices for the construction of one of our response variables, the weighted fractional rank deviation d.

Table 2 shows a summary of all variables with their descriptive statistics taking into account the sample weights. We encountered missing values in 20.5% of a total of 11654 registered children under the age of five. The final sample thus contains information on 9262 children. We observed most of the missing data, involving 17.5% of the children, for the height-for-age z-score. The remaining missing data resulted from education of the mother's partner, time to a water source, safe drinking water and satisfactory sanitation. We performed all data manipulations and subsequent analyses (see sections 6.2 and 6.3) using the statistical software package SAS, version 9.3 (SAS Institute, Cary, NC, USA).

*** Insert Table 2 around here ***

6.2 Regression Results

The six decompositions for comparison are based on the four regressions (6), (13), (21)–(22) and (26) that relate the degree of stunting h, the weighted fractional rank deviation d, both univariate as well as bivariate, and the composite variable v, to the set of nine explanatory variables. In each of the regressions, we used the sample weights to weigh the observations by sample area or household cluster. Typical for the bivariate regression model (21)–(22) is that the estimated coefficients coincide with those from fitting the two univariate models (6) and (13) using the same set of explanatory variables. In other words, the regression coefficients for h and d from a

univariate or bivariate estimation procedure are the same using the same set of explanatory variables. Table 3 shows the regression coefficients for h, d and v. It also contains the t- and F-statistics and significances for the univariate regressions. On the other hand, the bivariate regression provides bivariate test statistics and significances, which appear in Table 4.

*** Insert Tables 3 and 4 around here ***

The bivariate significances in Table 4 are based on Wilks' lambda or the likelihood ratio test that performs the same role as the t- or F-test in a univariate setting. Using Wilks' lambda, an approximation to F has been derived that closely fits its value (see, e.g., Tabachnick and Fidell, 2012). The F-test for the overall bivariate model shows that the model is highly significant. The F-tests for the bivariate response character confirm that the two response dimensions are appropriate and the F-tests for the individual explanatory variables show that all variables have a significant effect on the bivariate response except for time to a water source. The F-tests for the univariate regressions in Table 3 reveal that each of the univariate models is highly significant. The R^2 values of the models are small, however, especially for h and v. Also Wagstaff, Van Doorslaer and Watanabe (2003) observed small R^2 values for their stunting regressions (equal to 0.188 and 0.247) despite the fact that household consumption as a wealth-related variable was included in the models. Furthermore, the t-tests for the individual explanatory variables in the univariate regressions indicate that all variables are significant in at least one model.

Regarding the significant variables of the regression models in Table 3, we wish to highlight the following results. Apart from the absence of a wealth-related variable, our stunting model h resembles the stunting models presented by Wagstaff, Van Doorslaer and Watanabe (2003) and Van de Poel et al. (2007). Similar to Wagstaff, Van Doorslaer and Watanabe (2003), the t-statistics indicate that child's age is the most important determinant of malnutrition with coefficients for the linear and squared term describing an inverted u-shaped relationship. Other determinants of malnutrition are, in order of importance, the education of the mother and her partner, the residence type, the sex of the child and satisfactory sanitation. Furthermore, in line with our expectations, the most important effects on the fractional rank deviation d result from the residence type and safe drinking water, followed by the education of the mother's partner, satisfactory sanitation and the education of the mother. Except for the education of the mother, these variables are also important to explain v. In addition to that, v depends nonlinearly on child's age which is demonstrated by a normal u-shaped relationship.

6.3 Decomposition Results

Using either the 'product definition' in (1) or the 'covariance definition' in (5) of the Generalized Concentration Index, we obtained a value for the GC of -0.0136. Its negative sign reveals higher rates of child malnutrition amongst the poor or a socioeconomic inequality of malnutrition to the disadvantage of the poor.

Table 5 shows the absolute and percentage contributions of decompositions (I), (IIa), (IIb) and (IV) represented by (9), (15), (16) and (27). They are based on the univariate regressions for h, d and v. As opposed to decomposition (IV), decompositions (I), (IIa) and (IIb) are characterized by a residual term. These residuals have the same value of -0.0054 as a result from using the same set of nine explanatory variables in the regression models for h and d. Their percentage contribution equals 39.79%, which is substantial. Another result from comparing the four decompositions is the contrast between decompositions (I) and (IIb) without a constant term and decompositions (IIa) and (IV) with a constant term that exceeds more than three times the value of the GC in both decompositions. In addition to that, the rank-oriented decompositions (IIa) and (IIb) with and without a constant term, which are supposed to give similar results, are completely different. There is more similarity in decompositions (I) and (IIb) without a constant term and in decompositions (IIa) and (IV) with a constant term than in decompositions (IIa) and (IIb) that are based on the same regression model. The presence or absence of the constant term clearly dominates the decompositions. In decompositions (I) and (IIb), the residence type and the education of the mother and her partner are the most important contributors to socioeconomic inequality. Their contributions have the same sign as the GC, which means that they can be seen as factors which are responsible for the pro-poor character of socioeconomic inequality of malnutrition. In other words, living in towns and having parents with more years of education tend to be associated with less malnutrition. In decompositions (IIa) and (IV), on the other hand, the most important contributors are safe drinking water, followed by the education of the mother's partner and the residence type. However, the sign of their contributions is different from the sign of the GC, which means that they have a pro-rich effect on socioeconomic inequality of malnutrition. From a formal point of view, one might say that in these two cases the very large negative constant term forces most of the other terms to be positive.

*** Insert Table 5 around here ***

Next, Tables 6 and 7 contain the individual absolute contributions of

decompositions (IIIa) and (IIIb) represented by (24) and (25) where decomposition (IIIa) includes a constant term. They are based on the bivariate regression for h and d. Both decompositions examine the relationship between h and d, but also relate to the one-dimensional decompositions for h and d. In particular, the column and row totals of the contributions of decomposition (IIIa) in Table 6 relate to decompositions (I) and (IIa) and the column and row totals of the contributions of decomposition (IIIb) in Table 7 relate to decompositions (I) and (IIb). As a result, the residual term in decompositions (IIIa) and (IIIb) is the same as in decompositions (I), (IIa) and (IIb), amounting to -0.0054.

*** Insert Tables 6 to 7 around here ***

Tables 8 to 11 contain summary presentations of decompositions (IIIa) and (IIIb) showing the direct and combined or correlated contributions of the decompositions. Tables 8 and 9 contain the absolute contributions and Tables 10 and 11 the percentage contributions. Similar to decompositions (IIa) and (IV), the constant term in decomposition (IIIa) exceeds more than three times the value of the GC. It is offset by the direct contributions of the decomposition which sum to a large negative percentage. In contrast, the correlated contributions of the decomposition sum to a positive percentage that is about half the value of the GC. For decomposition (IIIb) without a constant term, the percentage totals from the direct and correlated contributions are all positive and therefore smaller in magnitude. In addition, the total of the correlated contributions is about twice as large as the total of the direct contributions.

The most important direct contributions to inequality in decomposition (IIIa) come from the same variables that determine inequality in decompositions (IIa) and (IV). Noting that in decomposition (IIIa) the large direct contributions of the linear and squared term of child's age balance each other out, these determinants are safe drinking water, followed by the education of the mother's partner and the residence type. Similarly, the direct contributions to inequality in decomposition (IIIb) correspond to those in decompositions (I) and (IIb) and come from the residence type and the education of the mother and her partner. As an overview, Figures 1 and 2 show the direct percentage contributions of decompositions (I), (IIb) and (IIIb) without a constant term and decompositions (IIa), (IIIa) and (IV) with a constant term.

*** Insert Tables 8 to 11 around here ***

*** Insert Figures 1 and 2 around here ***

7 Discussion

Both the theoretical framework and the empirical illustration reveal that there are many ways to generate regression-based decompositions of socioeconomic inequality of health, and that the results may be very different. We do not always find that the estimated individual contributions of the explanatory variables have the same sign, let alone the same magnitude. Three of the six decompositions have very large constant terms (by 'very large' we mean that these terms exceed by far the magnitude of the index itself), which seem difficult to interpret. The two super-decompositions involve a lot of correlation terms which on the whole may be more important than the direct contributions of the explanatory variables. Except for the decomposition based on the regression of the health-fractional rank deviation product, all decompositions have a residual term which may be very substantial in magnitude.

Both the one-dimensional rank-oriented approach and the two-dimensional simultaneous approach lead to two distinct decomposition formulas, one with and one without constant term. A comparison of the two sets of results shows huge differences. It is hard to understand why two equivalent starting points – the 'product definition' and the 'covariance definition' of the Generalized Concentration Index – produce so widely divergent outcomes. One might be tempted to conclude that any decomposition formula has a large amount of arbitrariness in it.

As far as the empirical aspects are concerned, perhaps we should have tried to increase the explanatory power of our regression equations by including more variables. Although this might improve the reliability of the empirical estimates, we doubt whether it would change much to the essence of our results. It seems highly unlikely that the constant term will fade away, that the correlation effects will become negligible, or that the residual term will disappear. Anyhow, additional empirical work (more years, more countries, more health variables, etc.) might be useful to try and discern a pattern in the various decompositions.

A special point to which we like to draw attention is the fact that we did not include socioeconomic status as an explanatory variable in our regression of health (or likewise health as an explanatory variable of socioeconomic status). Our motivation for doing so is that it seems unnatural to explain the correlation between health and socioeconomic status by including either of these variables. Oddly enough, this is often what happens in empirical work. E.g., in the study on child mortality in Iran by Hosseinpoor et al. (2006) the contribution of 'low economic status' was more than one third, and in the study on malnutrition in Ghana by Van de Poel et al. (2007) the 'wealth'

variable had a contribution of about 30%. In both cases, the socioeconomic variable had the highest contribution of all variables included in the model. In our view, this is not the proper way to decompose socioeconomic inequality.

8 Conclusion

In the past, research on the measurement of socioeconomic inequality of health has often been a copy of research on the measurement of income inequality. To a large extent, this has also been the case with regard to decomposition analysis. We believe that more caution should be exerted when adopting methods and results from one field to the other. The main reason is that bivariate inequality is of a different nature than univariate inequality. Moreover, the health variable taken into consideration is almost always not an unbounded ratio-scale variable such as income. What we have tried to show in this paper is that there are many ways of obtaining decomposition results for rank-dependent indicators of socioeconomic inequality of health, and that therefore decomposition results should not be taken for granted. In our opinion, one way to proceed from here is to use an axiomatic approach. This may be helpful first to identify which properties a good decomposition should have, and then to derive which decompositions possess the desired properties.

Appendix

In conventional matrix notation, the two regressions (21) and (22) can be written as $h = S\lambda + \psi$ and $d = S\pi + \chi$, where $S = [u, s_1, s_2, ..., s_k]$, and u is a vector of ones. The OLS estimates of λ and π are equal to $\widehat{\lambda} = (S'S)^{-1} S'h$ and $\widehat{\pi} = (S'S)^{-1} S'd$. This implies that the estimated errors are $\widehat{\psi} = h - S\widehat{\lambda} = Mh$ and $\widehat{\chi} = d - S\widehat{\pi} = Md$, where $M = I - S(S'S)^{-1} S'$. Matrix M is both symmetric and idempotent.

From $\widehat{\psi} = Mh$ it follows that $d'\widehat{\psi} = d'Mh$, and from $\widehat{\chi} = Md$ that $h'\widehat{\chi} = h'Md$. Since M is symmetric, we have $h'Md = h'M'd$, and of course $h'M'd = d'Mh$. Hence we have: $d'\widehat{\psi} = h'\widehat{\chi}$. Given that $Cov(d, \widehat{\psi}) = \frac{1}{n} d'\widehat{\psi}$ and $Cov(h, \widehat{\chi}) = \frac{1}{n} h'\widehat{\chi}$, it follows that $Cov(d, \widehat{\psi}) = Cov(h, \widehat{\chi})$. Moreover, we also have $\widehat{\psi}'\widehat{\chi} = (Mh)'(Md) = h'M'Md$. Since M is symmetric and idempotent, we know that $M'M = MM = M$, and so we find that $\widehat{\psi}'\widehat{\chi} = h'Md$. This means that $Cov(d, \widehat{\psi}) = Cov(h, \widehat{\chi}) = Cov(\widehat{\psi}, \widehat{\chi})$.

Table 1: Components of the six decompositions for comparison.

	Constant	Contribution of s_j	Correlation between s_j and s_g	Residual
I	–	$2\lambda_j Cov(s_j, d)$	–	$2Cov(\psi, d)$
IIa	$2\pi_0\mu_h$	$2\pi_j\left[Cov(h, s_j) + \mu_h\mu_{s_j}\right]$	–	$2Cov(h, \chi)$
IIb	–	$2\pi_j Cov(h, s_j)$	–	$2Cov(h, \chi)$
IIIa	$2\lambda_0\pi_0$	$2(\lambda_0\pi_j + \lambda_j\pi_0)\mu_{s_j} + 2\lambda_j\pi_j\left[Var(s_j) + (\mu_{s_j})^2\right]$	$2(\lambda_j\pi_g + \lambda_g\pi_j)\left[Cov(s_j, s_g) + \mu_{s_j}\mu_{s_g}\right]$	$2Cov(\psi, \chi)$
IIIb	–	$2\lambda_j\pi_j Var(s_j)$	$2(\lambda_j\pi_g + \lambda_g\pi_j)Cov(s_j, s_g)$	$2Cov(\psi, \chi)$
IV	$2\phi_0$	$2\phi_j\mu_{s_j}$	–	–

Table 2: Mean, standard deviation and description of all variables.

Variable	Mean	SD	Description
Degree of stunting	0.1252	0.2073	Height-for-age z-score (WHO) scaled to the interval [0,1]
			Degree of stunting > 0 if height-for-age z-score < -2 SD
Weighted fractional rank deviation	0	0.2952	Based on the wealth indices provided by DHS
Composite variable	-0.0068	0.0674	Degree of stunting $\times$ Weighted fractional rank deviation
Age of child	29.8571	17.8084	In months
Squared age of child	303.3724	270.6317	Term is mean-centered: $(\text{age of child} - 29.8571)^2$
Sex of child	0.5140	0.5110	Male (1), female (0)
Residence type	0.1237	0.3366	Urban (1), rural (0)
Education of mother	1.3446	2.8587	In years
Education of partner/husband	2.7439	3.8141	In years
Time to water source	0.8653	1.1756	In hours
Safe drinking water	0.4614	0.5097	Available (1), not available (0)
Satisfactory sanitation	0.1234	0.3362	Available (1), not available (0)

Table 3: Univariate regressions for the degree of stunting h, the weighted fractional rank deviation d and the composite variable $v \equiv hd$.

	h		d		v	
	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat
Constant	0.1307	22.37***	-0.1683	-25.07***	-0.0227	-11.88***
Age of child	0.0016	13.82***	0.0003	2.57*	-0.0001	-1.32
Squared age of child	-0.0001	-18.72***	0.0000	0.00	0.0000	2.09*
Sex of child	0.0135	3.34***	0.0067	1.43	-0.0003	-0.24
Residence type	-0.0257	-3.48***	0.2441	28.72***	0.0236	9.77***
Education of mother	-0.0036	-3.89***	0.0106	9.93***	-0.0004	-1.27
Education of partner/husband	-0.0030	-4.37***	0.0147	18.60***	0.0014	6.33***
Time to water source	-0.0003	-0.17	-0.0042	-2.06*	0.0005	0.79
Safe drinking water	0.0034	0.78	0.1298	25.91***	0.0175	12.30***
Satisfactory sanitation	-0.0171	-2.60**	0.1114	14.73***	0.0096	4.46***
F		85.82***		681.26***		72.47***
R^2	0.0770		0.3986		0.0659	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, else $p > 0.1$

Table 4: Bivariate analysis of variance of the degree of stunting h and the weighted fractional rank deviation d.

	Wilks' lambda	Bivariate F
Constant	0.8969	531.59***
Age of child	0.9785	101.43***
Squared age of child	0.9634	175.91***
Sex of child	0.9985	6.92***
Residence type	0.9179	413.94***
Education of mother	0.9883	54.65***
Education of partner/husband	0.9629	178.20***
Time to water source	0.9995	2.16
Safe drinking water	0.9318	338.38***
Satisfactory sanitation	0.9768	109.88***
Overall model	0.5605	345.05***
At least one distinct response	0.5605	345.05***
Two distinct responses	0.9346	80.90***

***$p < 0.001$, else $p > 0.1$

Table 5: Decompositions (I), (IIa), (IIb) and (IV).

	I		IIa		IIb		IV	
	value	%	value	%	value	%	value	%
Constant	–	–	-0.0422	309.98	–	–	-0.0454	333.72
Age of child	0.0001	-1.04	0.0029	-21.69	0.0004	-2.77	-0.0030	22.10
Squared age of child	0.0000	0.20	0.0000	-0.01	0.0000	0.01	0.0032	-23.38
Sex of child	0.0000	-0.27	0.0009	-6.62	0.0000	-0.32	-0.0003	2.41
Residence type	-0.0025	18.41	0.0046	-33.65	-0.0030	21.92	0.0058	-42.94
Education of mother	-0.0024	17.66	0.0022	-16.21	-0.0014	10.02	-0.0010	7.65
Education of partner/husband	-0.0028	20.58	0.0075	-55.19	-0.0026	18.92	0.0078	-57.29
Time to water source	0.0000	-0.12	-0.0010	7.00	0.0000	0.24	0.0008	-5.90
Safe drinking water	0.0004	-2.86	0.0143	-105.38	-0.0007	4.94	0.0162	-118.97
Satisfactory sanitation	-0.0010	7.66	0.0025	-18.03	-0.0010	7.26	0.0024	-17.41
Residual	-0.0054	39.79	-0.0054	39.79	-0.0054	39.79	–	–
Total	-0.0136	100.00	-0.0136	100.00	-0.0136	100.00	-0.0136	100.00

Table 6: Decomposition (IIIa) in relationship with decompositions (I) and (IIa).

	(1)	Age child	Squared age child	Sex child	Residence type	Education mother	Education partner	Time to water	Safe water	Satisfactory sanitation	χ	Total (I)
(1)	-0.0440	0.0027	0.0000	0.0009	0.0079	0.0037	0.0105	-0.0010	0.0157	0.0036	–	–
Age child	-0.0162	0.0013	0.0000	0.0003	0.0029	0.0013	0.0037	-0.0004	0.0058	0.0013	–	0.0001
Squared age child	0.0147	-0.0009	0.0000	-0.0003	-0.0027	-0.0013	-0.0036	0.0003	-0.0051	-0.0012	–	0.0000
Sex child	-0.0023	0.0001	0.0000	0.0001	0.0004	0.0002	0.0006	-0.0001	0.0008	0.0002	–	0.0000
Residence type	0.0011	-0.0001	0.0000	0.0000	-0.0016	-0.0003	-0.0006	0.0000	-0.0007	-0.0003	–	-0.0025
Education mother	0.0016	-0.0001	0.0000	0.0000	-0.0010	-0.0007	-0.0010	0.0000	-0.0009	-0.0003	–	-0.0024
Education partner	0.0028	-0.0002	0.0000	-0.0001	-0.0012	-0.0006	-0.0019	0.0001	-0.0013	-0.0004	–	-0.0028
Time to water	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	–	0.0000
Safe water	-0.0005	0.0000	0.0000	0.0000	0.0002	0.0001	0.0002	0.0000	0.0004	0.0001	–	0.0004
Satisfactory sanitation	0.0007	0.0000	0.0000	0.0000	-0.0004	-0.0001	-0.0003	0.0000	-0.0004	-0.0005	–	-0.0010
ψ	–	–	–	–	–	–	–	–	–	–	-0.0054	-0.0054
Total (IIa)	-0.0422	0.0029	0.0000	0.0009	0.0046	0.0022	0.0075	-0.0010	0.0143	0.0025	-0.0054	-0.0136

Table 7: Decomposition (IIIb) in relationship with decompositions (I) and (IIb).

	Age child	Squared age child	Sex child	Residence type	Education mother	Education partner	Time to water	Safe water	Satisfactory sanitation	χ	Total (I)
Age child	0.0003	0.0000	0.0000	0.0000	-0.0001	-0.0001	0.0000	0.0001	0.0000	–	0.0001
Squared age child	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0001	0.0000	0.0001	0.0000	–	0.0000
Sex child	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	–	0.0000
Residence type	0.0000	0.0000	0.0000	-0.0014	-0.0002	-0.0004	0.0000	-0.0004	-0.0002	–	-0.0025
Education mother	0.0000	0.0000	0.0000	-0.0007	-0.0006	-0.0006	0.0000	-0.0003	-0.0002	–	-0.0024
Education partner	0.0000	0.0000	0.0000	-0.0007	-0.0004	-0.0012	0.0000	-0.0003	-0.0002	–	-0.0028
Time to water	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	–	0.0000
Safe water	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0002	0.0000	–	0.0004
Satisfactory sanitation	0.0000	0.0000	0.0000	-0.0003	-0.0001	-0.0001	0.0000	-0.0001	-0.0004	–	-0.0010
ψ	–	–	–	–	–	–	–	–	–	-0.0054	-0.0054
Total (IIb)	0.0004	0.0000	0.0000	-0.0030	-0.0014	-0.0026	0.0000	-0.0007	-0.0010	-0.0054	-0.0136

Table 8: Decomposition (IIIa): values.

	Direct effect	Combined effect							
		Age child	Squared age child	Sex child	Residence type	Education mother	Education partner	Time to water	Safe water
Age child	-0.0122								
Squared age child	0.0147	-0.0009							
Sex child	-0.0014	0.0005	-0.0003						
Residence type	0.0074	0.0028	-0.0027	0.0004					
Education mother	0.0046	0.0012	-0.0013	0.0002	-0.0013				
Education partner	0.0114	0.0036	-0.0036	0.0005	-0.0018	-0.0017			
Time to water	-0.0009	-0.0004	0.0003	-0.0001	0.0000	0.0000	0.0000		
Safe water	0.0155	0.0059	-0.0051	0.0008	-0.0006	-0.0008	-0.0011	0.0000	
Satisfactory sanitation	0.0038	0.0013	-0.0012	0.0002	-0.0007	-0.0005	-0.0007	0.0000	-0.0003
Component total	0.0431	-0.0072							
(1)	-0.0440								
Residual	-0.0054								
Total	-0.0136								

Table 9: Decomposition (IIIb): values.

	Direct effect	Combined effect							
		Age child	Squared age child	Sex child	Residence type	Education mother	Education partner	Time to water	Safe water
Age child	0.0003								
Squared age child	0.0000	0.0000							
Sex child	0.0000	0.0000	0.0000						
Residence type	-0.0014	0.0000	0.0000	0.0000					
Education mother	-0.0006	-0.0001	0.0000	0.0000	-0.0009				
Education partner	-0.0012	-0.0001	-0.0001	0.0000	-0.0011	-0.0010			
Time to water	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		
Safe water	0.0002	0.0001	0.0001	0.0000	-0.0003	-0.0003	-0.0003	0.0000	
Satisfactory sanitation	-0.0004	0.0000	0.0000	0.0000	-0.0005	-0.0003	-0.0003	0.0000	-0.0001
Component total	-0.0030	-0.0052							
Residual	-0.0054								
Total	-0.0136								

Table 10: Decomposition (IIIa): percentages.

	Direct effect	Combined effect							
		Age child	Squared age child	Sex child	Residence type	Education mother	Education partner	Time to water	Safe water
Age child	89.73								
Squared age child	-107.86	6.38							
Sex child	9.97	-3.49	2.22						
Residence type	-54.48	-20.74	19.64	-2.90					
Education mother	-33.99	-8.81	9.32	-1.20	9.28				
Education partner	-83.87	-26.34	26.53	-3.71	13.35	12.21			
Time to water	6.38	2.64	-2.32	0.40	-0.02	-0.17	-0.27		
Safe water	-114.28	-43.13	37.50	-6.17	4.13	5.97	8.18	0.33	
Satisfactory sanitation	-28.17	-9.21	8.77	-1.33	5.08	3.43	5.43	-0.05	2.23
Component total	-316.56	53.15							
(1)	323.63								
Residual	39.79								
Total	100.00								

Table 11: Decomposition (IIIb): percentages.

	Direct effect	Combined effect							
		Age child	Squared age child	Sex child	Residence type	Education mother	Education partner	Time to water	Safe water
Age child	-2.48								
Squared age child	0.00	-0.19							
Sex child	-0.33	-0.02	0.03						
Residence type	10.01	0.15	0.30	0.03					
Education mother	4.40	0.54	0.19	0.01	6.45				
Education partner	9.00	0.92	0.75	0.00	7.82	7.61			
Time to water	-0.02	-0.01	0.04	0.01	0.03	0.04	0.02		
Safe water	-1.61	-0.46	-0.88	0.04	2.02	2.01	1.87	0.01	
Satisfactory sanitation	3.02	0.20	-0.03	-0.03	3.50	2.01	2.52	0.02	0.70
Component total	21.99	38.22							
Residual	39.79								
Total	100.00								

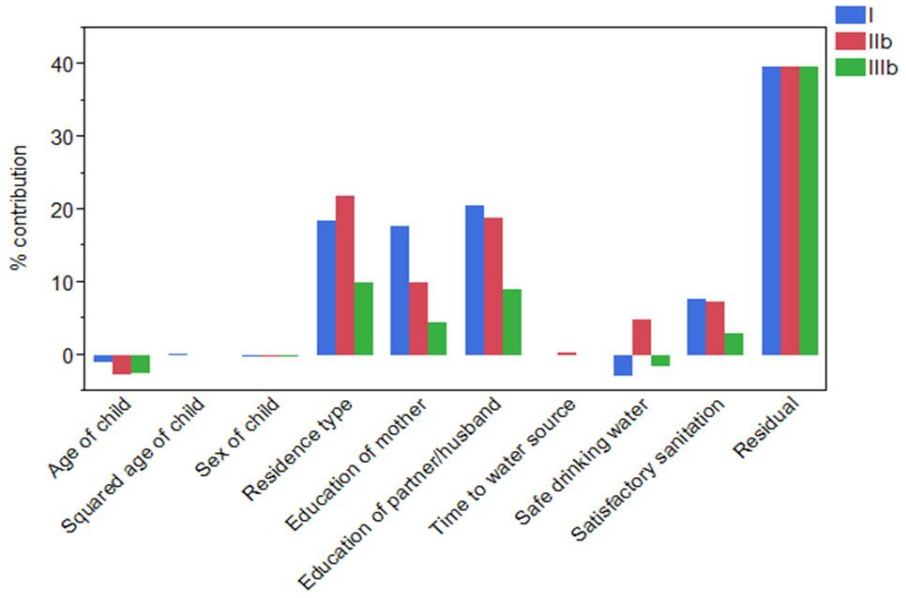

Figure 1: Contributions (%) from direct effects related to decompositions (I), (IIb) and (IIIb).

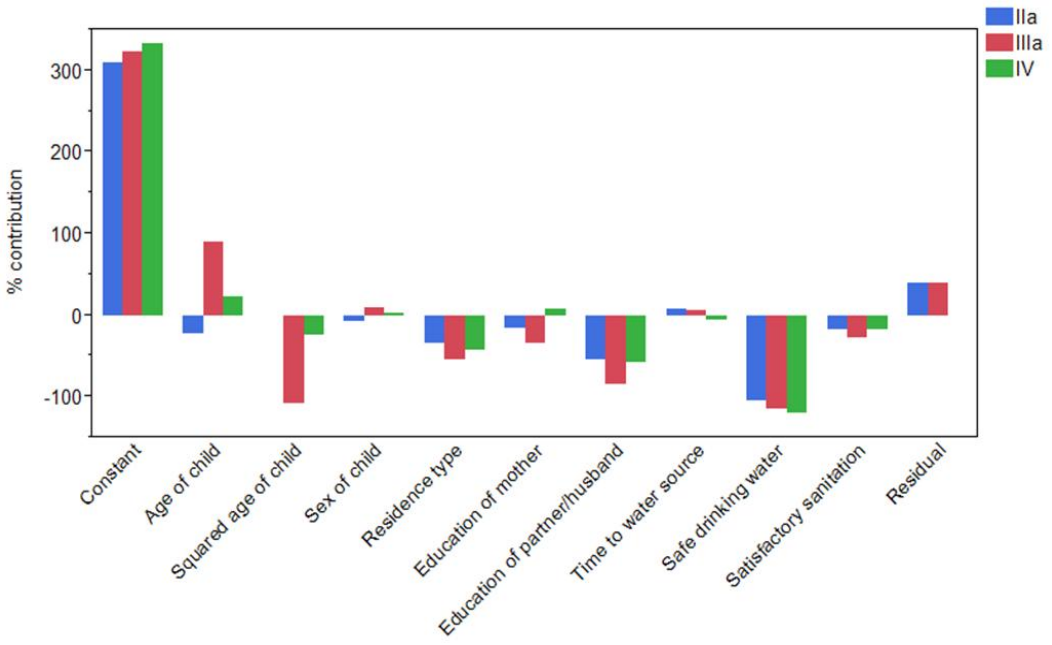

Figure 2: Contributions (%) from direct effects related to decompositions (IIa), (IIIa) and (IV).

References

[1] ABUL NAGA, Ramses H. and Pierre-Yves GEOFFARD (2006), "Decomposition of bivariate inequality indices by attributes", *Economics Letters*, 90(3), 362-367.

[2] BOURGUIGNON, François (1979), "Decomposable income inequality measures", *Econometrica*, 47(4), 901-920.

[3] CLARKE, Philip M., Ulf-G. GERDTHAM and Luke B. CONNELLY (2003), "A note on the decomposition of the health concentration index", *Health Economics*, 12(6), 511-516.

[4] COWELL, Frank A. and Carlo V. FIORIO (2011), "Inequality decompositions - a reconciliation", *Journal of Economic Inequality*, 9(4), 509-528.

[5] ERREYGERS, Guido (2009), "Correcting the concentration index", *Journal of Health Economics*, 28(2), 504-515.

[6] ERREYGERS, Guido and Tom 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(4), 685-694.

[7] FIELDS, Gary S. (2003), "Accounting for income inequality and its change: A new method, with application to the distribution of earnings in the United States", in: Solomon W. Polachek (ed.), *Worker Well-Being and Public Policy* (Research in Labor Economics, Volume 22), Bingley: Emerald Group Publishing Limited, pp. 1-38.

[8] GRAVELLE, Hugh (2003), "Measuring income related inequality in health: Standardisation and the partial concentration index", *Health Economics*, 12(10), 803-819.

[9] HOSSEINPOOR, Ahmad Reza, Eddy VAN DOORSLAER, Niko SPEYBROECK, Mohsen NAGHAVI, Kazem MOHAMMAD, Reza MAJDZADEH, Bahram DELAVAR, Hamidreza JAMSHIDI and Jeanette VEGA (2006), "Decomposing socioeconomic inequality in infant mortality in Iran", *International Journal of Epidemiology*, 35(5), 1211-1219.

[10] JONES, Andrew M. and Angel LÓPEZ NICOLÁS (2004), "Measurement and explanation of socioeconomic inequality in health with longitudinal data", *Health Economics*, 13(10), 1015-1030.

[11] MORDUCH, Jonathan and Terry SICULAR (2002), "Rethinking inequality decomposition, with evidence from rural China", *Economic Journal*, 112(476), 93-106.

[12] OAXACA, Ronald (1973), "Male-female wage differentials in urban labor markets", *International Economic Review*, 14(3), 693-709.

[13] O'DONNELL, Owen, Eddy VAN DOORSLAER and Adam WAGSTAFF (2006), "Decomposition of inequalities in health and health care", in: Andrew M. Jones (ed.), *The Elgar Companion to Health Economics*, Cheltenham: Edward Elgar, pp. 179-192.

[14] O'DONNELL, Owen, Eddy VAN DOORSLAER, Adam WAGSTAFF and Magnus LINDELOW (2008), *Analyzing Health Equity Using Household Survey Data. A Guide to Techniques and Their Implementation*, Washington, DC: World Bank.

[15] PODDER, Nripesh (1993), "The disaggregation of the Gini coefficient by factor components and its applications to Australia", *Review of Income and Wealth*, 39(1), 51-61.

[16] RAO, V. M. (1969), "Two decompositions of concentration ratio", *Journal of the Royal Statistical Society. Series A (General)*, 132(3), 418-425.

[17] SHORROCKS, Anthony F. (1980), "The class of additively decomposable inequality measures", *Econometrica*, 48(3), 613-625.

[18] SHORROCKS, Anthony F. (1982), "Inequality decomposition by factor components", *Econometrica*, 50(1), 193-211.

[19] SHORROCKS, Anthony F. (1984), "Inequality decomposition by population subgroups", *Econometrica*, 52(6), 1369-1385.

[20] TABACHNICK, Barbara G. and Linda S. FIDELL (2012), *Using Multivariate Statistics*, 6th Edition, Boston: Pearson Education.

[21] VAN DE POEL, Ellen, Ahmad Reza HOSSEINPOOR, Caroline JEHU-APPIAH, Jeanette VEGA and Niko SPEYBROECK (2007), "Malnutrition and the disproportional burden on the poor: The case of Ghana", *International Journal for Equity in Health*, 6(21) (doi:10.1186/1475-9276-6-21).

[22] VAN DOORSLAER, Eddy and Tom VAN OURTI (2011), "Inequality and inequity in health and health care", in: Sherry Glied and Peter

C. Smith (eds), *The Oxford Handbook of Health Economics*, Oxford: Oxford University Press, pp. 837-869.

[23] VAN OURTI, Tom, Eddy VAN DOORSLAER and Xander KOOLMAN (2009),"The effect of income growth and inequality on health inequality: Theory and empirical evidence from the European Panel", *Journal of Health Economics*, 28(3), 525-539.

[24] WAGSTAFF, Adam (2005), "The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality", *Health Economics*, 14(4), 429-32.

[25] WAGSTAFF, Adam, Eddy VAN DOORSLAER and Naoko WATANABE (2003), "On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam", *Journal of Econometrics*, 112(1), 207-223.