DRAFT

This paper is a draft submission to

Inequality—Measurement, trends, impacts, and policies

5–6 September 2014   Helsinki, Finland

A unified structural equation modeling approach for the decomposition of rank-dependent indicators of socioeconomic inequality of health

Roselinde Kessels and Guido Erreygers

July 23, 2014

Abstract

In this paper we present a unified structural equation modeling (SEM) framework for the regression-based decomposition of rank-dependent indicators of socioeconomic inequality of health and compare it with simple ordinary least squares (OLS) regression. The SEM framework forms the basis for a proper use of the most prominent one- and two-dimensional decompositions and provides an argument for using the bivariate multiple regression model for two-dimensional decomposition. Within the SEM framework, the two-dimensional decomposition integrates the feedback mechanism between health and socioeconomic status and allows for different sets of determinants of these variables. We illustrate the SEM approach and its outperformance to OLS using data from the Ethiopia 2011 Demographic and Health Survey (DHS).

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

1 Introduction

The dominant approach with regard to the measurement of socioeconomic inequality of health consists of using rank-dependent indicators. They are called rank-dependent because they can be expressed as weighted averages of individual health levels, with the weights determined by the ranks of individuals in the socioeconomic distribution. Indices of this type allow us to find out whether there is pro-rich or pro-poor bias in the health distribution: positive values indicate that people who are relatively well-off in socioeconomic terms tend to have better health than those who are less well-off, and negative values the opposite. The standard health Concentration Index (Wagstaff, Paci and Van Doorslaer, 1991) is undoubtedly the most popular rank-dependent index. There is now also an increasing literature on the decomposition of the Concentration Index using various econometric techniques (see, e.g., Wagstaff, Van Doorslaer and Watanabe, 2003; Van Ourti, Van Doorslaer and Koolman, 2009; Allanson and Petrie, 2013). An overview of recent contributions on the measurement and decomposition of socioeconomic inequality of health can be found in Van Doorslaer and Van Ourti (2011) and Van Ourti, Erreygers and Clarke (2014).

Compared to indicators of income inequality or health inequality, which measure the degree of inequality within a given univariate distribution for income or health, indicators of socioeconomic inequality of health are bivariate in nature because they measure the degree of *correlation* between health and socioeconomic status. To explain the degree of correlation between these two variables rather than the degree of inequality in one variable, Erreygers and Kessels (2013) proposed a set of two-dimensional decompositions that investigate both variables simultaneously. The most salient of these decompositions is based on the bivariate multiple regression model explaining health and socioeconomic status simultaneously. This decomposition captures not only the direct contributions of the explanatory variables in the regressions, but also their combined or correlated contributions.

However, two criticisms may be made of the two-dimensional decomposition analysis based on the bivariate multiple regression model. The first is that the bivariate multiple regression model uses the same set of variables to explain both health and socioeconomic status, which may not be the most appropriate assumption given that the determinants of health and socioeconomic status need not be the same. Related to this, the second criticism is that socioeconomic status is not included as an explanatory variable in the regression of health, and health not included as an explanatory variable in the regression of socioeconomic status. The existence of a reciprocal relationship might be examined since health is potentially both a cause and a conse-

quence of socioeconomic status (O'Donnell, Van Doorslaer and Van Ourti, 2014). In the literature on the decomposition of socioeconomic inequality of health, several empirical studies (see, e.g., Wagstaff, Van Doorslaer and Watanabe, 2003; Hosseinpoor et al., 2006; Van de Poel et al., 2007; Doherty, Walsh and O'Neill, 2014) have investigated the impact of socioeconomic status on health, reporting evidence that socioeconomic status is an important determinant of health.

The main objective of Erreygers and Kessels (2013) was to compare the two-dimensional decomposition to the one-dimensional decompositions that are based on regressions of only one of the two variables under consideration. Therefore, they used the same set of explanatory variables in all regressions, which are all estimated using ordinary least squares (OLS). Moreover, for the one-dimensional decompositions, they argued that including either of the variables as an explanatory variable in the single regressions distorts the explanation of the correlation between health and socioeconomic status. It is then as if the variable in question is treated both as a dependent and as an independent variable. As a result, for the two-dimensional decomposition, the bivariate multiple regression modeling framework was chosen which includes neither health nor socioeconomic status as an explanatory variable.

To bridge the gap between empirical observations and modeling practice, we propose a flexible modeling approach for the decomposition of socioeconomic inequality of health that makes use of a structural or simultaneous equation model (SEM). The model allows for different sets of determinants of health and socioeconomic status as well as for the inclusion of socioeconomic status as an explanatory variable in the regression of health and health as an explanatory variable in the regression of socioeconomic status. The model produces consistent estimates of the regression coefficients using a two-step generalized method of moments (GMM) estimation procedure that includes instrumental variables. Although such a modeling approach has been hinted at before (Wagstaff, Van Doorslaer and Watanabe, 2003: 214, n.12), our paper is the first to adopt it in practice.

The outline of the remainder of the paper is as follows. Section 2 reviews the Generalized health Concentration Index and the various concepts it embraces. Section 3 gives an overview of the most important one- and two-dimensional decompositions based on OLS regression. In Section 4, we present a unified SEM approach for regression-based decomposition analysis and show how it fits with the existing decompositions. By means of an empirical analysis of child malnutrition in Ethiopia in Section 5, we illustrate the proper use of the one- and two-dimensional decompositions within the OLS and SEM regression framework. Finally, Section 6 summarizes the paper and discusses the main outcomes.

2 Generalized health Concentration Index

We consider a population of n individuals for which the health level of individual i, denoted as h_i, is 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 $\mu_h = \frac{1}{n} \sum_{i=1}^{n} h_i$.

Erreygers and Van Ourti (2011) pointed out that the use of the health Concentration Index is apropos 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 these indices belong to the family of rank-dependent indices: they can be expressed as weighted sums of health levels with the weights determined by socioeconomic ranks. The socioeconomic rank of individual i is determined by his/her position according to the variable chosen to measure socioeconomic well-being, e.g. income. Let the value of this variable for individual i be y_i. Then the natural number $r_i(y)$, or more simply r_i, measures the position of individual i in the rank-order according to variable y, with the rank $r_i = 1$ assigned to the person who is least well-off, and the rank $r_i = n$ assigned to the person who is most well-off. 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 is defined as:

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

The standard health Concentration Index as well as the indices introduced by Wagstaff (2005) and Erreygers (2009) are simple functions of GC. We can rewrite the formula for GC using a well-known relationship 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 $\mu_d = 0$, the value for GC can also be computed as:

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

Erreygers and Kessels (2013) used both (1) and (2) to generate decompositions of the Generalized Concentration Index. Some of these decompositions

have a constant term. Because it is problematic to give a meaningful interpretation to the constant term, the most attractive decompositions are those without a constant term. In the next section, we review these decompositions, two of which are one-dimensional and one is two-dimensional. In the one-dimensional decompositions, either the health variable or the fractional rank deviation variable is subject to a regression, whereas in the two-dimensional decomposition, both variables are subject to a regression. The regression approach used is simply OLS.

3 One- and two-dimensional decompositions using OLS regression

3.1 The health-oriented decomposition

The health-oriented decomposition, introduced by Wagstaff, Van Doorslaer and Watanabe (2003), has been the first and most well-known regression-based decomposition. It starts from the linear regression model describing the relationship between the health variable h and a number of explanatory variables $x_1, x_2, ..., x_k$:

$$h_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \varepsilon_i \tag{3}$$

where ε_i is an error term. Substituting the right-hand side of this model for h_i in the 'product definition' of the GC in (1) and working out the result, we obtain the health-oriented decomposition, henceforth referred to as decomposition (I):

$$GC = 2 \sum_{j=1}^{k} \beta_j Cov(x_j, d) + 2Cov(\varepsilon, d) \tag{4}$$

This decomposition has a deterministic component consisting of a sum of k contributions, one for each explanatory variable, and a residual component.

As argued by Erreygers and Kessels (2013), it is misleading to include the fractional rank deviation variable d in the OLS regression for h in decomposition (I), or any proxy variable strongly correlated with d such as income or consumption. In that case, the residual component will be zero, or close to zero, suggesting that we have explained all or most of the variation in the Generalized Concentration Index. This result is, however, merely an artefact from the OLS regression-based approach of decomposition (I). Consider,

for example, the simple case where the variable d is the only explanatory variable of h, i.e. $x_1 = d$. Since the OLS estimate of β_1 is then equal to $Cov(h, d)/Var(d)$, it follows that the deterministic component of decomposition (I) is identical to GC and therefore the residual component equal to zero. However, in this case, we have explained nothing at all. We are just treating the fractional rank deviation variable d both as a dependent and as an independent variable.

Even though empirical work suggests that the socioeconomic variable is an important predictor for *health* (see, e.g., Wagstaff, Van Doorslaer and Watanabe, 2003; Hosseinpoor et al., 2006; Van de Poel et al., 2007; Doherty, Walsh and O'Neill, 2014), the OLS regression-based methodology of decomposition (I) does not provide the right framework to use this result for the explanation of *socioeconomic inequality of health*. To bridge the gap between the empirical result and the regression-based decomposition methodology, we propose using a SEM approach (see Section 4) that unifies these contrasting themes.

3.2   A rank-oriented decomposition

Erreygers and Kessels (2013) introduced a rank-oriented decomposition that relies on a linear regression model for the fractional rank deviations. Assuming that the variables $z_1, z_2, ..., z_q$ are the relevant variables to explain the socioeconomic ranks, this model is given by

$$d_i = \gamma_0 + \gamma_1 z_{1,i} + \gamma_2 z_{2,i} + ... + \gamma_q z_{q,i} + \xi_i \tag{5}$$

where ξ_i is an error term. Substituting the right-hand side of this model for d_i in the 'covariance definition' of the GC in (2) and working out the result, we arrive at the rank-oriented decomposition, henceforth referred to as decomposition (II):

$$GC = 2 \sum_{g=1}^{q} \gamma_g Cov(h, z_g) + 2Cov(h, \xi) \tag{6}$$

Decomposition (II) has a similar structure as decomposition (I) because it decomposes the Generalized Concentration Index into a sum of q explained contributions, with each of these equal to a covariance weighted by a regression coefficient, and a residual or unexplained component, which is also a covariance. In line with proper practice to exclude the socioeconomic variable d in the OLS regression for h in decomposition (I), Erreygers and Kessels (2013) also advise against the inclusion of h in the OLS regression for d because it would artificially result in a zero residual covariance in decomposition

(II). In this case, too, in order to make room for a possible effect of health on socioeconomic status in the framework of decomposition (II), we recommend using a SEM procedure that describes the feedback mechanism between these two variables (see Section 4).

3.3   A two-dimensional simultaneous decomposition

To attend to the bivariate nature of the Generalized Concentration Index, Erreygers and Kessels (2013) proposed a set of two-dimensional decompositions that investigate the health levels h and the fractional rank deviations d simultaneously. The most salient of these decompositions is based on the bivariate multiple regression model explaining both variables simultaneously. Typical for the bivariate multiple regression is that a common set of p variables $s_1, s_2, ..., s_p$ is used to explain h and d. The bivariate multiple regression has the following form:

$$h_i = \lambda_0 + \lambda_1 s_{1,i} + \lambda_2 s_{2,i} + ... + \lambda_p s_{p,i} + \psi_i \tag{7}$$

$$d_i = \pi_0 + \pi_1 s_{1,i} + \pi_2 s_{2,i} + ... + \pi_p s_{p,i} + \chi_i \tag{8}$$

where ψ_i and χ_i are error terms. It is assumed that $\mu_\psi = \mu_\chi = 0$ and that the $2p$ covariances $Cov(s_j, \chi)$ and $Cov(\psi, s_j)$ are zero.

Applying the 'covariance definition' of the GC in (2) to the bivariate multiple regression model leads to the simultaneous decomposition, henceforth referred to as decomposition (III):

$$
\begin{aligned}
GC &= 2\sum_{j=1}^{p} \lambda_j \pi_j Var(s_j) + 2\sum_{j=1}^{p} \sum_{g=j+1}^{p} (\lambda_j \pi_g + \lambda_g \pi_j) Cov(s_j, s_g) \\
&\quad + 2Cov(\psi, \chi)
\end{aligned}
\tag{9}
$$

It consists of p single-variable terms $\lambda_j \pi_j Var(s_j)$ which capture the direct effect of the p explanatory variables, $\frac{p(p-1)}{2}$ two-variable terms $(\lambda_j \pi_g + \lambda_g \pi_j) \times Cov(s_j, s_g)$ which capture the correlation structure between the explanatory variables, and a residual component which is proportional to the covariance between the two error terms.

In the next section, we show that the simultaneous decomposition based on the bivariate multiple regression model is also the decomposition that we obtain from applying a SEM regression approach.

4   A unified structural equation modeling approach for decomposition analysis

Perhaps the most pertinent critique of the bivariate multiple regression model as a basis for two-dimensional decomposition is the one that questions the assumption that the same set of p variables explains both the health variable h and the fractional rank deviation d. The challenge rests on the grounds that the determinants of health and socioeconomic status need not be the same. Moreover, the bivariate multiple regression model seems inflexible in the sense that it does not include h as a predictor in the equation for d and d as a predictor in the equation for h. Empirical evidence has shown, however, that health is largely influenced by socioeconomic status. It might also be the case that socioeconomic status is influenced by health, implying that both variables influence one another reciprocally.

To overcome the criticisms of the bivariate multiple regression model, we propose the specification of a structural or simultaneous equation model (SEM) (see, e.g., Greene, 2011, chapter 10; Verbeek, 2012, chapter 5) which allows for different sets of predictors for h and d as well as the addition of d as a predictor in the equation for h and of h as a predictor in the equation for d. These structural equations are meant to represent causal relationships among the variables in the model.

We assume that the variables $x_1, x_2, ..., x_k$ with $x_k = d$ are the relevant variables in the equation for h and $z_1, z_2, ..., z_q$ with $z_q = h$ are the relevant variables in the equation for d. We then have the following structural model of two equations (SEM):

$$h_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{j,i} + \beta_k d_i + \varepsilon_i \tag{10}$$

$$d_i = \gamma_0 + \sum_{g=1}^{q-1} \gamma_g z_{g,i} + \gamma_q h_i + \xi_i \tag{11}$$

In the SEM, the variables h and d are assumed endogenous or jointly determined by the system of simultaneous equations. The random error terms ε and ξ affect both h and d (which is made clear by rewriting (11) in terms of h), suggesting a correlation between each of the endogenous variables and each of the random error terms. The remainder of the variables in the SEM are assumed exogenous or determined outside the system.

Because of the endogeneity of the variables h and d, OLS regression cannot be relied upon to produce consistent estimates of the parameters

of the equations. Instead, a generalized method of moments (GMM) estimation procedure using instrumental variable (IV) estimation or two-stage least squares (2SLS) is needed to consistently estimate all parameters of the SEM (Hansen, 1982). This requires the introduction of at least one instrumental variable or instrument for each equation. An instrument for an equation is strongly correlated with the right-hand side endogenous variable of that equation but uncorrelated with the equation's error term. Moreover, an instrument does not have a direct effect on the response variable, and thus it does not belong on the right-hand side of the equation as an explanatory variable. It is therefore only a tool or instrument to solve the endogeneity problem, hence the name. Using an efficient GMM estimator, a necessary condition for identification of the two-equation SEM is that each equation has at least one exogenous variable that is not present in the other equation.

Once the SEM is estimated, Equation (10) can be used as input for decomposition (I) and Equation (11) as input for decomposition (II). In this way, by using an efficient GMM estimation procedure instead of OLS, the contribution of d in decomposition (I) and of h in decomposition (II) is duly measured.

Substituting the right-hand side of (11) for d_i in (10) and the right-hand side of (10) for h_i in (11), we obtain:

$$h_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{j,i} + \beta_k \left[\gamma_0 + \sum_{g=1}^{q-1} \gamma_g z_{g,i} + \gamma_q h_i + \xi_i \right] + \varepsilon_i$$

$$d_i = \gamma_0 + \sum_{g=1}^{q-1} \gamma_g z_{g,i} + \gamma_q \left[\beta_0 + \sum_{j=1}^{k-1} \beta_j x_{j,i} + \beta_k d_i + \varepsilon_i \right] + \xi_i$$

Rearranging terms and assuming that $\beta_k \gamma_q \neq 1$, we arrive at the following reformulation of the model, which is called the reduced form of the SEM:

$$h_i = \frac{\beta_0 + \beta_k \gamma_0}{1 - \beta_k \gamma_q} + \sum_{j=1}^{k-1} \frac{\beta_j}{1 - \beta_k \gamma_q} x_{j,i} + \sum_{g=1}^{q-1} \frac{\beta_k \gamma_g}{1 - \beta_k \gamma_q} z_{g,i} + \frac{\varepsilon_i + \beta_k \xi_i}{1 - \beta_k \gamma_q} \quad (12)$$

$$d_i = \frac{\gamma_0 + \beta_0 \gamma_q}{1 - \beta_k \gamma_q} + \sum_{j=1}^{k-1} \frac{\beta_j \gamma_q}{1 - \beta_k \gamma_q} x_{j,i} + \sum_{g=1}^{q-1} \frac{\gamma_g}{1 - \beta_k \gamma_q} z_{g,i} + \frac{\xi_i + \gamma_q \varepsilon_i}{1 - \beta_k \gamma_q} \quad (13)$$

The reduced-form equations express each endogenous variable, h and d, in terms of the exogenous variables, $x_1, x_2, ..., x_{k-1}$ and $z_1, z_2, ..., z_{q-1}$, and the intercept, plus an error term. If variable x_{j*} is equal to variable z_{g*} – nothing excludes this case – then the coefficient of the variable in question in (12) will be $(\beta_{j*} + \beta_k \gamma_{g*})/(1 - \beta_k \gamma_q)$, and in (13) $(\beta_{j*} \gamma_q + \gamma_{g*})/(1 - \beta_k \gamma_q)$.

Like the bivariate multiple regression model (7)–(8), the reduced form of the SEM in (12)–(13) is characterized by the same set of explanatory variables, which we note as $s_1, s_2, ..., s_p$. Equations (12) and (13) can then be simplified as:

$$h_i = \lambda_0 + \lambda_1 s_{1,i} + \lambda_2 s_{2,i} + ... + \lambda_p s_{p,i} + \psi_i \tag{14}$$

$$d_i = \pi_0 + \pi_1 s_{1,i} + \pi_2 s_{2,i} + ... + \pi_p s_{p,i} + \chi_i \tag{15}$$

The parameters $\lambda_0, \lambda_1, ..., \lambda_p$ and $\pi_0, \pi_1, ..., \pi_p$ in (14)–(15) are called reduced-form parameters. The error terms ψ_i and χ_i are called reduced-form errors.

The reduced-form equations (14)–(15) are equivalent to the bivariate multiple regression model (7)–(8), and can be directly estimated by OLS since the right-hand side variables are exogenous and uncorrelated with the random errors ψ_i and χ_i. This shows that using a SEM regression approach, we end up with decomposition (III) based on the bivariate multiple regression model. Within the SEM framework, this decomposition incorporates the feedback mechanism between the variables h and d which are allowed to depend on different sets of predictors. As a result, the above analysis answers the criticisms of the bivariate multiple regression model and the resulting decomposition (III).

In our empirical illustration described in the next section, we show that a SEM regression analysis forms the basis for a proper use of decompositions (I), (II) and (III).

5 An empirical illustration

5.1 Data description

For comparison the data are the same as those used by Erreygers and Kessels (2013). They come from the 2011 Demographic and Health Survey (DHS) of Ethiopia and are confined to children under the age of five.

The response variables in decompositions (I), (II) and (III) are the health variable h and the fractional rank deviation d. The health variable h is actually an ill-health variable: the degree of stunting or malnutrition. It is defined on the unit interval $[0, 1]$ and provides information on the depth of malnutrition with children. It is measured using 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. The new WHO child growth population was chosen as reference

population. The degree of stunting is stated 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, 44% of the children in the dataset are stunted. The fractional rank deviation d was obtained by ranking the children's households according to their wealth status using the wealth indices constructed by DHS from a principal component analysis on all household living conditions and assets. In the computation sample weights were taken into account so that, in effect, the variable d stands for the weighted fractional rank deviation.

The set of explanatory variables is the same as the one used by Erreygers and Kessels (2013) except for the variable time to water source, which turned out to be insignificant in their decomposition analyses. The variables are age and sex of the child, education of the mother and her partner or husband, urban or rural residence, access to safe drinking water and satisfactory sanitation. In addition to that, child's age is specified nonlinearly in the regression models using a squared term, which is mean-centered to remove multicollinearity with the linear term. Furthermore, safe drinking water and satisfactory sanitation are defined along the lines proposed by the WHO and UNICEF. Safe drinking water includes the following sources of water supply: 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. Satisfactory sanitation includes the following sanitation infrastructure: 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.

Table 1 shows a summary of all variables with their descriptive statistics taking into account the sample weights. The data contain information on 9262 children under the age of five. The value for the GC equals -0.0136 using either the 'product definition' in (1) or the 'covariance definition' in (2). Its negative sign reveals higher rates of child malnutrition amongst the poor or a socioeconomic inequality of malnutrition to the disadvantage of the poor. In the next two sections, we apply the various approaches described in this paper to compute decompositions (I), (II) and (III), which we express in percentages. We first discuss the decomposition results from using OLS regression, and then those from using a SEM approach. We performed all regression analyses using the econometric software package EViews 8.

11

5.2 Decomposition results using OLS regression

5.2.1 Decompositions (I) and (II)

The results for decompositions (I) and (II) depend on the specification of the OLS regression model used. For decomposition (I), an important comparison to study is that between excluding and including the weighted fractional rank deviation d in the regression for the degree of stunting h. For decomposition (II), we carry out a similar analysis comparing the results from excluding and including h in the regression for d. Table 2 contains the coefficients for the two sets of regressions for h and d as well as the t- and F-statistics and significances. We corrected standard errors for heteroskedasticity by using White's heteroskedasticity-consistent standard errors.

Using OLS regression the t-statistics indicate that the variables d and h are highly significant in the regressions for h and d, respectively. In other words, h is very much influenced by d, and vice versa, d is very much influenced by h. Furthermore, the regression results for h are largely affected when d is included as a regressor, whereas the regression results for d do not seem to differ much when h is included. When d is excluded in the regression for h, all variables are significant at the 5% level except for safe drinking water. However, when d is included in the regression for h, two more variables besides safe drinking water turn out to be insignificant, namely residence type and satisfactory sanitation. Also, education of the mother and her partner have become less significant when including d as a regressor. Consequently, it seems that d is attributing some of the explanatory variation of these variables in the regression for h. In contrast, when h is included in the regression for d, only child's age is affected in the sense that it has become more significant. Although we argue that the OLS framework is not the right methodology to estimate the regression models with d and h as regressors, because of the correlated nature of the cross-sectional data, we suspect that the regression model for h including d makes more sense than the regression model for d including h.

Using the two regressions for h, excluding and including d, we computed two versions of decomposition (I) and using the two regressions for d, excluding and including h, we computed two versions of decomposition (II). The percentage contributions of these decompositions are shown in Table 3 and visualized in Figure 1. An important observation is that decomposition (I) has a zero residual component when d is included and decomposition (II) has a zero residual component when h is included. Also, the contributions for d in decomposition (I) and for h in decomposition (II) are by far the largest, being 66.08% and 43.08%, and seem to capture all residual variation

on top of their real contributions when compared to the large residual value of 39.76% of decomposition (I) excluding d and of decomposition (II) excluding h. As discussed in Section 3.1, this result is an artefact of including either socioeconomic status or health as a variable in the decompositions that aim to explain the correlation between these variables.

One might thus inadvertently conclude that the contributions of d and h are very large in decompositions (I) and (II). However, for decomposition (I), the contribution of d exceeds the residual term from excluding d by a factor of 1.66, whereas for decomposition (II), the contribution of h is about the same as the residual term from excluding h. Compared to these residual terms, the contribution of the socioeconomic variable in decomposition (I) may be real and large, but not as large as 66.08%, whereas the contribution of the health variable in decomposition (II) may not be real. Also, similar to the regression results, when d is included in decomposition (I), the contributions of most other variables are smaller in absolute magnitude than when d is excluded. In contrast, when h is included in decomposition (II), the contributions of the other variables seem largely unaffected.

5.2.2 Decomposition (III)

We computed decomposition (III) starting from the bivariate multiple regression model, the coefficients of which are the same as those from the univariate regressions for h excluding d and for d excluding h, shown in Table 2. Table 4 contains the individual percentage contributions of decomposition (III). As indicated by Erreygers and Kessels (2013), the column and row totals of the contributions of decomposition (III) relate to decompositions (I) and (II) from the regressions for h excluding d and for d excluding h. The contribution of the residual term in decomposition (III) is therefore the same as in decompositions (I) and (II), equating to 39.76%. Table 5 contains a summary presentation of decomposition (III) showing the direct and combined or correlated percentage contributions. Similar to the results of Erreygers and Kessels (2013), the total of the combined or correlated contributions is almost twice as large as the total of the direct contributions. As a comparison, Figure 2 contains the direct percentage contributions of decomposition (III) as well as the contributions from decomposition (I) excluding d and from decomposition (II) excluding h.

Assuming that all explanatory variables in the bivariate multiple regression model are the exogenous variables in a two-equation SEM for the estimation of h and d, decomposition (III) takes into account the mutual dependency between h and d and thus captures the net or reduced effects of the explaining variables upon both h and d. In the next section, we further

discuss the relevant decompositions obtained from using a SEM approach.

5.3 Decomposition results using a SEM approach

5.3.1 SEM estimation

A first step in a SEM regression analysis for the estimation of h and d as endogenous variables in Equations (10)–(11) is to define the exogenous variables for each equation as well as the instrumental variables for GMM analysis. Looking at the OLS regression results in Table 2, we learn that the variables child's age, both its linear and squared term, and sex of child are important predictors for h, but not for d, whereas the variables residence type, safe drinking water and satisfactory sanitation are important predictors for d, but not for h. We have therefore removed the variables residence type, safe drinking water and satisfactory sanitation in the equation for h, and used residence type and satisfactory sanitation as instruments for d. We did not include safe drinking water as an instrument because we obtained a more powerful GMM analysis by not considering this variable. Similarly, we have removed the variables child's age, both its linear and squared term, and sex of child in the equation for d, and used all three terms as instruments for h. For each equation in the SEM, we then have at least one exogenous variable that is not present in the other equation, so that our system is identified.

We estimated the SEM in (10)–(11) using a feasible efficient two-step GMM procedure for robust covariance estimation in the presence of heteroskedasticity (White, 1982) using EViews 8. This procedure is also known as two-stage instrumental variables (2SIV) or heteroskedastic two-stage least squares (H2SLS). Table 6 contains the GMM regression coefficients for the two-equation SEM as well as the OLS regression coefficients for comparison. Regarding the GMM analysis, Table 6 includes the t-, Hansen's J- and Cragg-Donald F-statistics and significances. Hansen's J-statistic has a χ^2-distribution under the null hypothesis that the instruments for an equation in the SEM are valid. The J-statistics for the two equations in the SEM are not significant at the 5% level so that we conclude that all our instruments are valid. The Cragg-Donald F-statistic is used to test for weak instruments or instruments that are not highly correlated with an equation's right-hand side endogenous variable. The Cragg-Donald F-statistics for the two equations in the SEM are highly significant meaning that the instruments for each equation are strong.

Using GMM regression the t-statistics indicate that the health variable h is largely influenced by the weighted fractional rank deviation d, but there is no feedback or two-way influence in the sense that the weighted fractional

rank deviation d is not affected by h. This result is different from the result from OLS regression shown in Table 6 where h is highly significant in the regression for d (see also the discussion in Section 5.2.1). Furthermore, most exogenous variables in the GMM analysis of the SEM are significant at the 5% level. Only in the GMM regression for h education of the mother's partner is insignificant and education of the mother is significant at the 10% level.

5.3.2   Decompositions

Because the GMM analysis has shown that d has a significant impact on h, but not vice versa, we use the GMM regression for h of Table 6 as input for decomposition (I) and we can simply use the OLS regression for d excluding h of Table 2 as input for decomposition (II). We refer to Section 5.2.1 where we computed decomposition (II) based on this OLS regression. Note that whether or not we include the variables child's age, both its linear and squared term, and sex of child in the OLS regression for d does not make much difference for decomposition (II). Table 7 contains the percentage contributions of decomposition (I) based on the SEM's equation for h in (10). We report the contributions using the GMM regression coefficients of Table 6 as well as the OLS regression coefficients for comparison. Figure 3 visualizes the two sets of contributions of decomposition (I). Note that, to compute the contribution of the weighted fractional rank deviation d using the GMM regression coefficients, we did not include d itself, but the *predicted* value of d resulting from the OLS regression of d on all exogenous and instrumental variables of the SEM's equation for h.

Table 7 and Figure 3 show that decomposition (I) using GMM regression has a large residual component of 38.11%, which is of the same size as that of decomposition (I) excluding d using OLS regression. Furthermore, the contribution of d is much lower and more realistic using GMM instead of OLS. It reduced from 68.45% using OLS to 42.62% using GMM, which is, however, still a substantial percentage. Also, the contributions of the variables education of the mother and her partner have lowered to the same extent, approximately by a factor of 0.6, using GMM instead of OLS.

Lastly, regarding decomposition (III), whether we use the bivariate multiple regression model or the SEM regression approach, we end up with the same decomposition (III) which we discussed in Section 5.2.2.

15

6 Summary and discussion

Empirical research has provided evidence that socioeconomic status, represented by a wealth- or income-related variable, is an important determinant of health. Vice versa, it is likely that health is an important determinant of socioeconomic status. However, to explain socioeconomic inequality of health, or the *correlation* between health and socioeconomic status, neither of the variables, health or socioeconomic status, can be used as an explanatory variable in an OLS regression-based decomposition approach, because we would then explain the bivariate dependent variable by one of its univariate components, which is meaningless. To unify the potentially bidirectional relationship between health and socioeconomic status with the regression-based decomposition methodology, we recommend using a structural or simultaneous equation model (SEM) which captures the feedback mechanism between health and socioeconomic status using a system of equations for these variables, which are assumed endogenous. More specifically, this two-equation model allows the inclusion of socioeconomic status as an explanatory variable for health and health as an explanatory variable for socioeconomic status, while providing consistent estimates using a two-step generalized method of moments (GMM) estimation procedure. Also, it allows the specification of different sets of determinants of health and socioeconomic status.

The SEM for the estimation of health and socioeconomic status can easily be transformed into a bivariate multiple regression model for these variables, which, in the SEM framework, is also called the reduced form of the SEM. The SEM's exogenous variables are the explanatory variables in each equation of this model format. We can then simply apply OLS to estimate the bivariate multiple regression model and use the regression coefficients as input for the two-dimensional simultaneous decomposition introduced by Erreygers and Kessels (2013). As such, we have shown that this decomposition takes into account the mutual dependency between health and socioeconomic status and captures the reduced effects of the explaining variables upon health and socioeconomic status.

In case one wishes to use one of the main one-dimensional decompositions, the health-oriented decomposition proposed by Wagstaff, Van Doorslaer and Watanabe (2003) or the rank-oriented decomposition (without a constant term) proposed by Erreygers and Kessels (2013), a GMM analysis of the SEM is required for a proper application of these decompositions. However, if the GMM regressions indicate that socioeconomic status is insignificant in the equation for health or health in the equation for socioeconomic status, we can resort again to an OLS regression analysis of the particular equation(s) after having removed the insignificant endogenous variable(s). By all means

we advise against applying OLS to each of the initial SEM equations.

In our empirical illustration, the GMM analysis of the SEM confirms previous findings that health is largely influenced by socioeconomic status, but the opposite relationship appears not to hold. In the GMM analysis the effect of socioeconomic status on health is, however, indirect and measured by the instrumental variables residence type and satisfactory sanitation. We used the GMM regression coefficients of the health equation in the health-oriented decomposition and the OLS regression coefficients of the socioeconomic status equation (without an explanatory health variable) in the rank-oriented decomposition. We recommend such modeling practice when computing one-dimensional decompositions. The contribution of socioeconomic status in the health-oriented decomposition turns out to be 42.62%, which is substantial and by far the largest. This contribution is, however, indirect and measured by the variables residence type and satisfactory sanitation. The residual term is not zero compared to using OLS regression coefficients, but amounts to 38.11%, which is about the same size as the residual term from decompositions based on OLS regressions without health and socioeconomic status as explanatory variables.

Furthermore, we computed the two-dimensional simultaneous decomposition based on the bivariate multiple regression model since this model is equivalent to the reduced form of the SEM. The total of the combined or correlated contributions in this decomposition is almost twice as large as the total of the direct contributions and the residual term amounts to 39.76%. In all, we can conclude that the SEM provides a unified modeling framework for correctly applying the one- and two-dimensional decompositions and we therefore recommend it as a starting basis for decomposition analysis.

Finally, the SEM proposed in this paper is an observed-variables SEM because the endogenous variables health and socioeconomic status are observed or measured. A potentially interesting topic for further research would be to construct a SEM where the endogenous variables are not observed, but latent (Jöreskog, 1973). In this respect, it may prove useful to make the socioeconomic variable, when assumed latent, directly dependent on a series of wealth-related variables. Also, instead of transforming the socioeconomic levels into ranks, another avenue would be to use the socioeconomic levels as such, and to construct a level-dependent index of socioeconomic inequality of health as proposed by Erreygers and Kessels (2014) to which a SEM regression-based decomposition analysis can easily be applied.

Table 1: 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
Age of child	29.8571	17.8084	In months
Squared age of child	303.3724	270.6317	Term is mean-centered: (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
Safe drinking water	0.4614	0.5097	Available (1), not available (0)
Satisfactory sanitation	0.1234	0.3362	Available (1), not available (0)

Table 2: OLS regressions for the degree of stunting h and the weighted fractional rank deviation d, where d has been excluded and included in the regression for h and h has been excluded and included in the regression for d.

| | h | | | | d | | | |
| | excluding d | | including d | | excluding h | | including h | |
	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat
Constant	0.1305	15.80***	0.1212	14.16***	-0.1720	-18.54***	-0.1627	-16.99***
Age of child	0.0016	10.94***	0.0016	11.10***	0.0003	1.82◇	0.0005	2.41*
Squared age of child	-0.0001	-13.49***	-0.0001	-13.49***	0.0000	0.02	0.0000	-0.80
Sex of child	0.0135	2.30*	0.0139	2.36*	0.0065	0.98	0.0074	1.12
Residence type	-0.0255	-2.18*	-0.0122	-1.02	0.2470	22.30***	0.2452	21.84***
Education of mother	-0.0036	-3.43***	-0.0030	-2.87**	0.0106	8.06***	0.0103	7.87***
Education of partner/husband	-0.0030	-3.31***	-0.0022	-2.38*	0.0146	13.49***	0.0144	13.28***
Safe drinking water	0.0033	0.53	0.0103	1.60	0.1289	18.12***	0.1291	18.18***
Satisfactory sanitation	-0.0170	-2.03*	-0.0110	-1.28	0.1118	12.12***	0.1106	11.96***
d	—	—	-0.0539	-4.19***	—	—	—	—
h	—	—	—	—	—	—	-0.0712	-4.17***
F	96.55***		90.11***		765.62***		687.07***	
R^2	0.0770		0.0806		0.3983		0.4006	

$^{◇}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

19

Table 3: Percentage contributions from decomposition (I), excluding and including d, and decomposition (II), excluding and including h, using the OLS regressions of Table 2.

	I		II	
	excluding d	including d	excluding h	including h
Age of child	-1.04	-1.05	-2.79	-3.72
Squared age of child	0.20	0.20	0.04	-1.55
Sex of child	-0.27	-0.28	-0.31	-0.36
Residence type	18.26	8.74	22.18	22.02
Education of mother	17.65	14.86	10.03	9.79
Education of partner/husband	20.60	15.18	18.88	18.61
Safe drinking water	-2.81	-8.66	4.91	4.91
Satisfactory sanitation	7.65	4.94	7.30	7.22
d	—	66.08	—	—
h	—	—	—	43.08
Residual	39.76	0	39.76	0
Total	100.00	100.00	100.00	100.00

Table 4: Percentage contributions from decomposition (III) in relationship with decompositions (I) and (II) using the OLS regressions for h excluding d and for d excluding h of Table 2.

	Age child	Squared age child	Sex child	Residence type	Education mother	Education partner	Safe water	Satisfactory sanitation	χ	Total (I)
Age child	-2.49	0.00	-0.02	0.16	0.59	0.96	-0.46	0.21	–	-1.04
Squared age child	-0.19	0.04	0.03	0.31	0.19	0.75	-0.88	-0.03	–	0.20
Sex child	-0.01	0.00	-0.32	0.03	0.02	0.00	0.04	-0.03	–	-0.27
Residence type	0.00	0.00	0.00	10.05	1.51	2.64	2.65	1.42	–	18.26
Education mother	-0.04	0.00	0.00	4.99	4.41	4.75	2.16	1.39	–	17.65
Education partner	-0.04	0.00	0.00	5.22	2.86	8.99	2.13	1.45	–	20.60
Safe water	0.00	0.00	0.00	-0.66	-0.16	-0.27	-1.57	-0.14	–	-2.81
Satisfactory sanitation	-0.01	0.00	0.00	2.09	0.62	1.08	0.83	3.03	–	7.65
ψ	–	–	–	–	–	–	–	–	39.76	39.76
Total (II)	-2.79	0.04	-0.31	22.18	10.03	18.88	4.91	7.30	39.76	100.00

Table 5: Percentage direct and combined contributions from decomposition (III) using the individual contributions of Table 4.

	Direct effect	Combined effect						
		Age child	Squared age child	Sex child	Residence type	Education mother	Education partner	Safe water
Age child	-2.49							
Squared age child	0.04	-0.19						
Sex child	-0.32	-0.02	0.03					
Residence type	10.05	0.15	0.31	0.03				
Education mother	4.41	0.54	0.19	0.01	6.50			
Education partner	8.99	0.92	0.75	0.00	7.86	7.60		
Safe water	-1.57	-0.46	-0.88	0.04	1.99	2.00	1.86	
Satisfactory sanitation	3.03	0.21	-0.03	-0.03	3.51	2.01	2.52	0.69
Component total	22.13				38.11			
Residual	39.76							
Total					100.00			

Table 6: GMM regressions (and OLS regressions for comparison) for the structural equation model (SEM) which includes the degree of stunting h and the weighted fractional rank deviation d as endogenous variables.

| | h | | | | d | | | |
| | GMM | | OLS | | GMM | | OLS | |
	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat
Constant	0.1187	13.52***	0.1240	15.32***	-0.1700	-16.01***	-0.1493	-26.15***
Age of child	0.0017	11.18***	0.0016	11.13***	—	—	—	—
Squared age of child	-0.0001	-13.48***	-0.0001	-13.55***	—	—	—	—
Sex of child	0.0143	2.41*	0.0138	2.34*	—	—	—	—
Residence type	—	—	—	—	0.2502	22.55***	0.2457	21.94***
Education of mother	-0.0022	-1.81$^\diamond$	-0.0033	-3.36***	0.0108	8.01***	0.0102	7.80***
Education of partner/husband	-0.0014	-1.27	-0.0024	-2.63***	0.0148	13.37***	0.0144	13.21***
Safe drinking water	—	—	—	—	0.1288	17.96***	0.1296	18.23***
Satisfactory sanitation	—	—	—	—	0.1132	12.17***	0.1108	11.97***
d	-0.0987	-3.46***	-0.0559	-4.67***	—	—	—	—
h	—	—	—	—	0.0826	1.25	-0.0621	-3.73***
R^2	0.0767		0.0796		0.3895		0.3996	
J	0.42		—		2.69		—	
Cragg-Donald F	917.43***		—		194.31***		—	

$^\diamond p < 0.1$, $^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$

Table 7: Percentage contributions from decomposition (I) using the SEM's equation for h and the GMM and OLS regression coefficients of Table 6.

	GMM	OLS
Age of child	-1.06	-1.06
Squared age of child	0.20	0.20
Sex of child	-0.29	-0.28
Residence type	—	—
Education of mother	10.75	16.21
Education of partner/husband	9.67	16.47
Safe drinking water	—	—
Satisfactory sanitation	—	—
d	42.62	68.45
h	—	—
Residual	38.11	0
Total	100.00	100.00

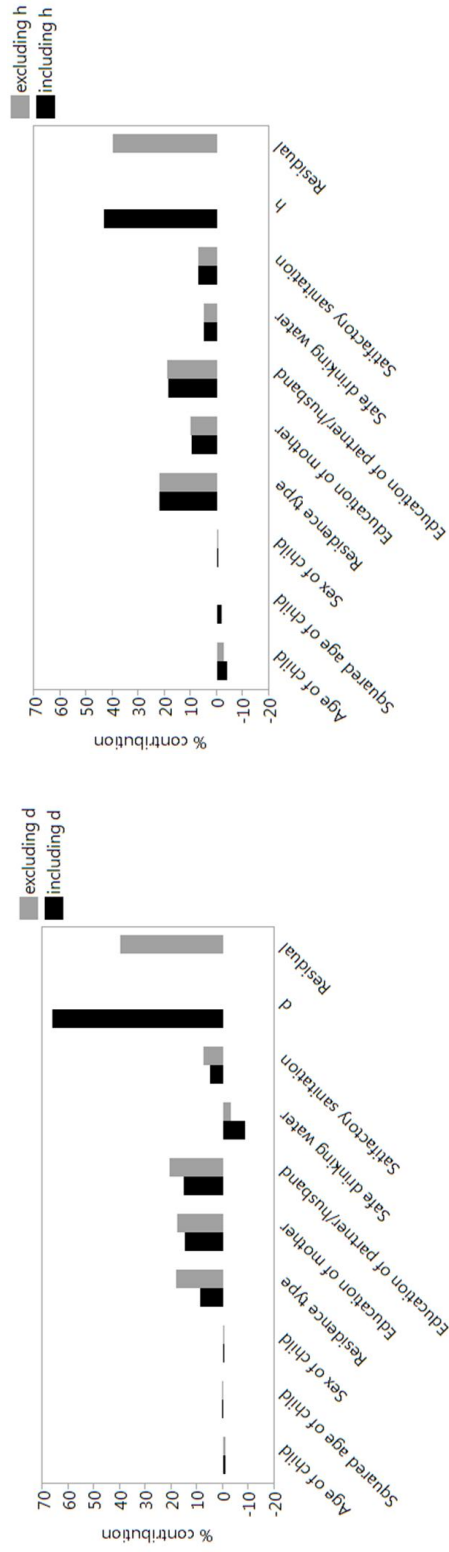(a) Decomposition I

(b) Decomposition II

Figure 1: Percentage contributions from decomposition (I), excluding and including d, and decomposition (II), excluding and including h, using the OLS regressions of Table 2.

Figure 2: Percentage contributions from direct effects related to decompositions (I), (II) and (III) using the OLS regressions for h excluding d and for d excluding h of Table 2.

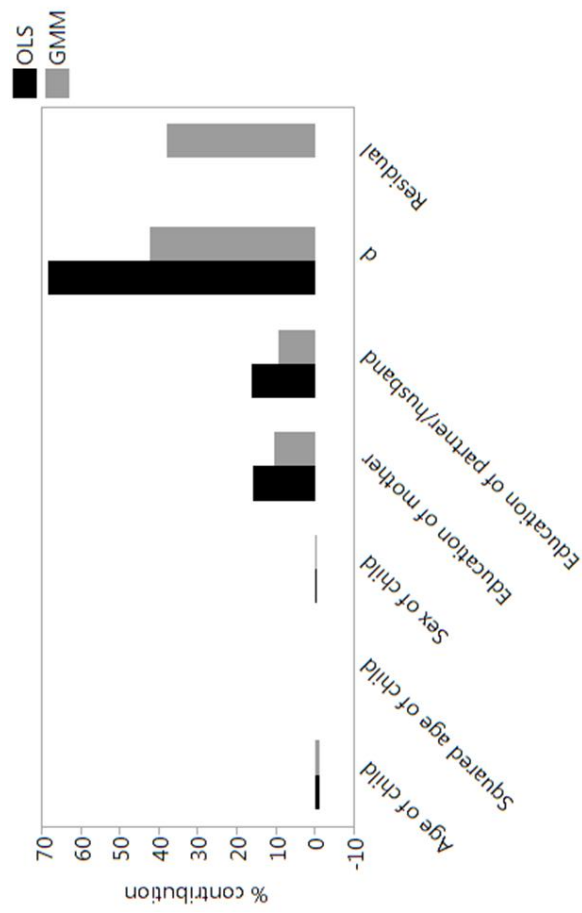Figure 3: Percentage contributions from decomposition (I) using the SEM's equation for h and the GMM and OLS regression coefficients of Table 6.

References

[1] ALLANSON, Paul and Dennis PETRIE (2013), "Longitudinal methods to investigate the role of health determinants in the dynamics of income-related health inequality", *Journal of Health Economics*, 32(5), 922-937.

[2] DOHERTY, Edel, Brendan WALSH and Ciaran O'NEILL (2014), "Decomposing socioeconomic inequality in child vaccination: Results from Ireland", *Vaccine*, 32(27), 3438-3444.

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

[4] ERREYGERS, Guido and Roselinde KESSELS (2013), "Regression-based decompositions of rank-dependent indicators of socioeconomic inequality of health", in Pedro Rosa Dias and Owen O'Donnell (eds.), *Health and Inequality* (Research on Economic Inequality, Volume 21, Chapter 9), London: Emerald Group Publishing Limited, pp. 227-259.

[5] ERREYGERS, Guido and Roselinde KESSELS (2014), "Socioeconomic status and health: A new approach to the measurement of bivariate inequality", Research report, Faculty of Applied Economics, Universiteit Antwerpen, in progress.

[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] GREENE, William H. (2011), *Econometric Analyis*, 7th Edition, New York: Macmillan.

[8] HANSEN, Lars P. (1982), "Large sample properties of generalized method of moments estimators", *Econometrica*, 50(4), 1029-1054.

[9] HOSSEINPOOR, Ahmad Reza, Eddy VAN DOORSLAER, Niko SPEY-BROECK, Mohsen NAGHAVI, Kazem MOHAMMAD, Reza MA-JDZADEH, 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] JORESKOG, Karl G. (1973), "A general method for estimating a linear structural equation system", in Arthur S. Goldberger and Otis Dudley

Duncan (eds.), *Structural Equation Models in the Social Sciences*, New York: Seminar Press, pp. 85-112.

[11] O'DONNELL, Owen, Eddy VAN DOORSLAER and Tom VAN OURTI (2014), "Health and inequality", in Anthony B. Atkinson and François J. Bourguignon (eds.), *Handbook of Income Distribution*, Volume 2B, Chapter 18, Amsterdam: Elsevier.

[12] 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).

[13] VAN DOORSLAER, Eddy and Tom VAN OURTI (2011), "Measuring 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, Chapter 35, pp. 837-869.

[14] VAN OURTI, Tom, Guido ERREYGERS and Philip CLARKE (2014), "Measuring equality and equity in health and health care", in Anthony J. Culyer (ed.), *Encyclopedia of Health Economics*, San Diego: Elsevier, pp. 234-239.

[15] 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.

[16] VERBEEK, Marno (2012), *A Guide to Modern Econometrics*, 4th Edition, West Sussex, U.K.: John Wiley & Sons.

[17] 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), 429-432.

[18] WAGSTAFF, Adam, Pierella PACI and Eddy VAN DOORSLAER (1991), "On the measurement of inequalities in health", *Social Science and Medicine*, 33(5), 545-557.

[19] 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.

[20] WHITE, Halbert (1982), "Instrumental variables regression with independent observations", *Econometrica*, 50(2), 483-499.