A Unified Structural Equation Modeling Approach for the Decomposition of Rank-Dependent Indicators of Socioeconomic Inequality of Health

Roselinde Kessels & Guido Erreygers

Friday 5 September 2014

UNITED NATIONS UNIVERSITY
UNU-WIDER

Universiteit Antwerpen

Socioeconomic Inequality of Health

- Deals with *two dimensions*: socioeconomic status (SES) and health
- Widely measured by *rank-dependent indicators*: they measure SES by the ranks which individuals occupy in the socioeconomic distribution, and health (or ill-health) by the levels of the health variable under consideration
- Most well-known indicator is the *Concentration Index (CI)*, which has two versions: the *relative* or *standard CI* and the *absolute* or *generalized CI*

Relative and Generalized Concentration Curves

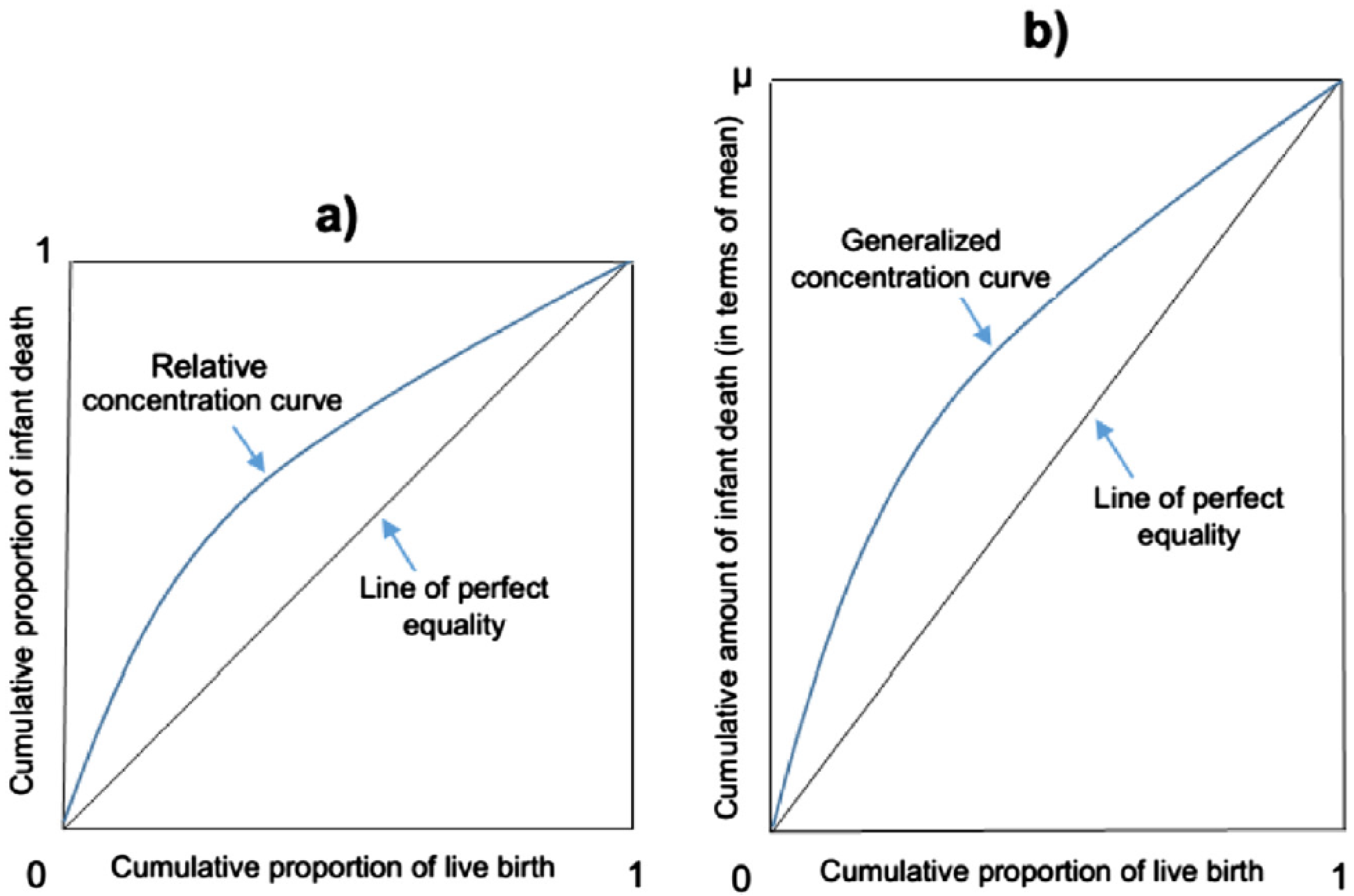


Fig. 1. Relative and generalized concentration curves.

Aim of the Paper

- To provide the right framework for a regression-based decomposition analysis to explain the *generalized CI (GC)*, which measures the degree of *correlation* between health and SES
- We show that a *structural equation modeling (SEM)* framework forms the basis for proper use of existing decompositions
- We highlight the one-dimensional decompositions where either health or SES is subject to a regression and the most salient two-dimensional simultaneous decomposition proposed by Erreygers and Kessels (2013)

Basic Notations

- Population of n individuals (1, 2, …, n)
- Health variable h, individual health levels $h_1, h_2, \ldots, h_n$
  - Ratio-scale (nonnegative) or cardinal (with finite lower bound)
- SES variable y, individual levels $y_1, y_2, \ldots, y_n$
- SES rank variable $r = r(y)$, individual ranks $r_1, r_2, \ldots, r_n$
  - Least well-off individual has rank 1, most well-off rank n; average $\mu_r = (n + 1)/2$
  - Fractional ranks $f_i \equiv 1/n \times (r_i - ½)$; average $\mu_f = ½$
  - Fractional rank deviations $d_i \equiv f_i - \mu_f$; average $\mu_d = 0$

Generalized Health Concentration Index (GC)

- Product definition

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

- Covariance definition

$$GC = 2Cov(h, d)$$

Health-Oriented Decomposition

- Introduced by Wagstaff, Van Doorslaer & Watanabe (2003)
- Starting point is the regression of health *h*

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

- Using the product definition of the *GC*, it follows that

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

- This leads to decomposition (I)

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

Rank-Oriented Decomposition

- Introduced by Erreygers & Kessels (2013)
- Starting point is the regression of the fractional rank deviation variable *d*

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

- Using the covariance definition of the *GC* results in decomposition (II)

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

Two-Dimensional Simultaneous Decomposition

- Introduced by Erreygers & Kessels (2013)
- Starting point is the bivariate multiple regression model explaining *h* and *d* simultaneously

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

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

- Using the covariance definition of the *GC* results in decomposition (III)

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

Criticisms of the OLS Regression Models

1. The bivariate multiple regression model uses the same set of variables to explain both *h* and *d*
   - This may not be the most appropriate assumption given that the determinants of *h* and *d* need not be the same
2. In all our OLS models, the variable *d* is not included as an explanatory variable in the regression for *h*, and *h* is not included as an explanatory variable in the regression for *d*
   - The existence of a reciprocal relationship might be examined since health is potentially both a cause and a consequence of SES (O'Donnell, Van Doorslaer & Van Ourti, 2014)

OLS Regressions for *h* and *d* with *d* and *h* as Predictors

- It is misleading to include *d* (or any proxy variable strongly correlated with *d* such as income or consumption) in the OLS regression for *h* in decomposition (I) and *h* in the OLS regression for *d* in decomposition (II)
- The residual component of the decompositions will be zero, or close to zero, which is an artificial result
- E.g.: the simple regression of *h* on $x_1 = d$ has an OLS estimate of β_1 equal to $Cov(h,d) / Var(d)$ so that

$$\begin{aligned} GC &= 2\frac{Cov(h,d)}{Var(d)}Cov(d,d) + 2Cov(\varepsilon,d) \\ &= 2Cov(h,d) + 0 \end{aligned}$$

OLS Regression for *h* with SES as Predictor

- Frequently applied in decomposition (I) (e.g., Wagstaff, Van Doorslaer & Watanabe, 2003; Hosseinpoor et al., 2006; Van de Poel et al., 2007; Doherty, Walsh & O'Neill, 2014)
- The contribution of SES to the *GC* in decomposition (I) has been artificially large (~ 30%)
- However, it has been shown that SES is an important determinant of health
- How to combine this empirical result with the regression-based decomposition methodology?

SEM Approach

- Starting point is the two-equation SEM

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

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

  - The variables *h* and *d* are assumed endogenous
  - To consistently estimate all parameters, estimation occurs through generalized method of moments (GMM) using instrumental variables (IV)

SEM Approach

- Substituting for *d* and *h* on the right-hand side of the equations yields

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

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

SEM Approach

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

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

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

SEM Approach

- The reduced-form equations are equivalent to the bivariate multiple regression model; they include the same set of explanatory variables, and can be directly estimated by OLS

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

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

SEM Approach

- Results in decomposition (III) based on the bivariate multiple regression model
- Thus, decomposition (III) integrates the feedback mechanism between the variables *h* and *d* which are allowed to depend on different sets of predictors
- This refutes the two criticisms of the bivariate multiple regression model and the resulting decomposition (III)

Empirical Illustration: Data

- We look at stunting of children below the age of five in Ethiopia
- The data come from the latest round (2011) of the Demographic and Health Survey (DHS) of Ethiopia
- Our dataset contains 9262 children
- Stunting (malnutrition) is defined as having a low height-for-age *z*-score (i.e. *z*-score < -2 SD from median height-for-age of reference population)
- We converted stunting into a continuous bounded variable ("0" = *z*-score ≥ -2 SD; "1" = *z*-score = -6 SD)
- We selected a set of 8 variables (exogenous & instruments)
- We performed weighted regressions, using the sample weights of the DHS dataset

Descriptive Statistics

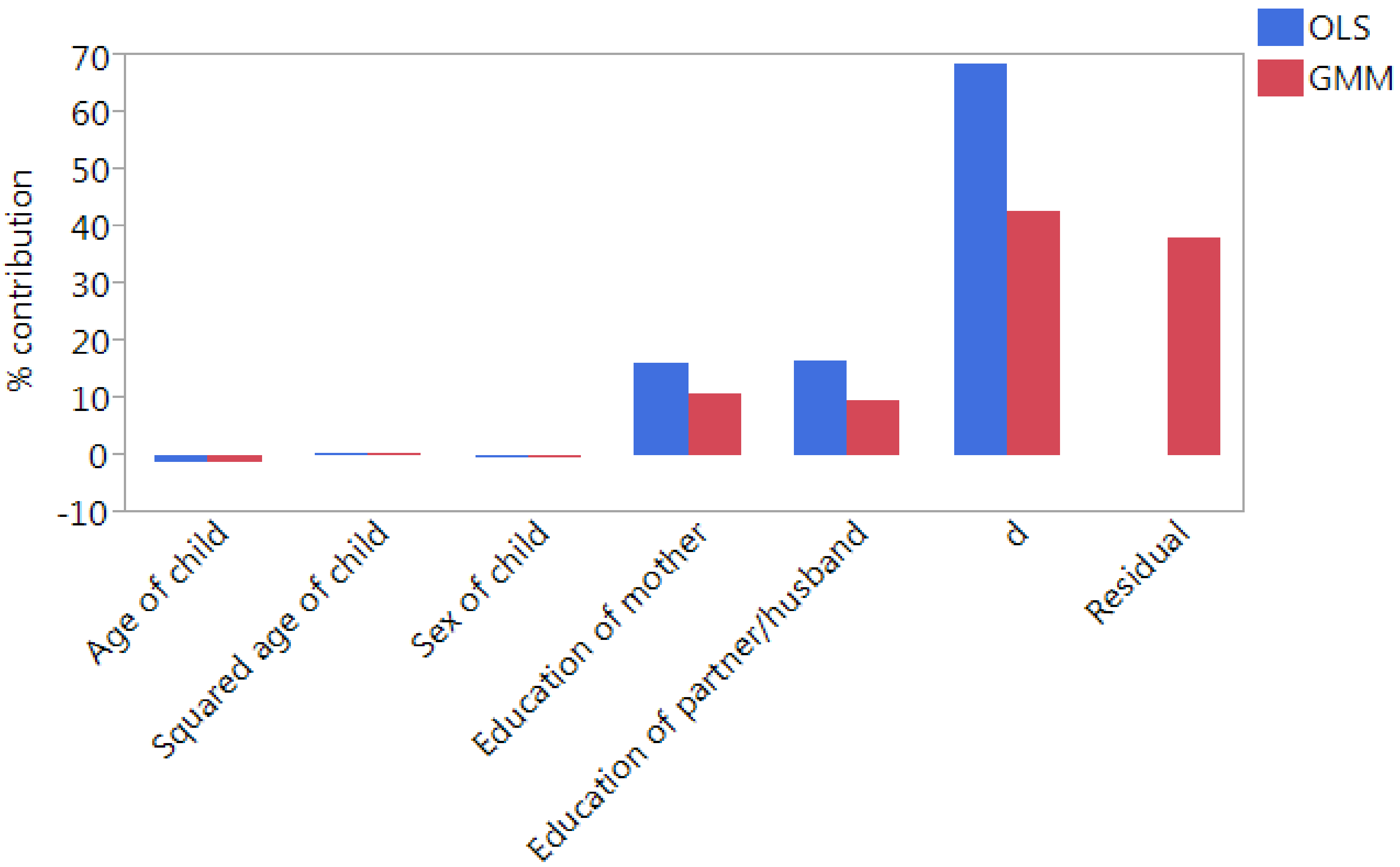
Variable	Mean	SD	Description
Degree of stunting	0.1252	0.2073	Height-for-age z-score (WHO) scaled to the interval [0,1]
			Degree of stunting > 0 if height-for-age z-score < -2 SD
Weighted fractional rank deviation	0	0.2952	Based on the wealth indices provided by DHS
Age of child	29.8571	17.8084	In months
Squared age of child	303.3724	270.6317	Term is mean-centered: $(\text{age of child} - 29.8571)^2$
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
Safe drinking water	0.4614	0.5097	Available (1), not available (0)
Satisfactory sanitation	0.1234	0.3362	Available (1), not available (0)

$GC = -0.0136$

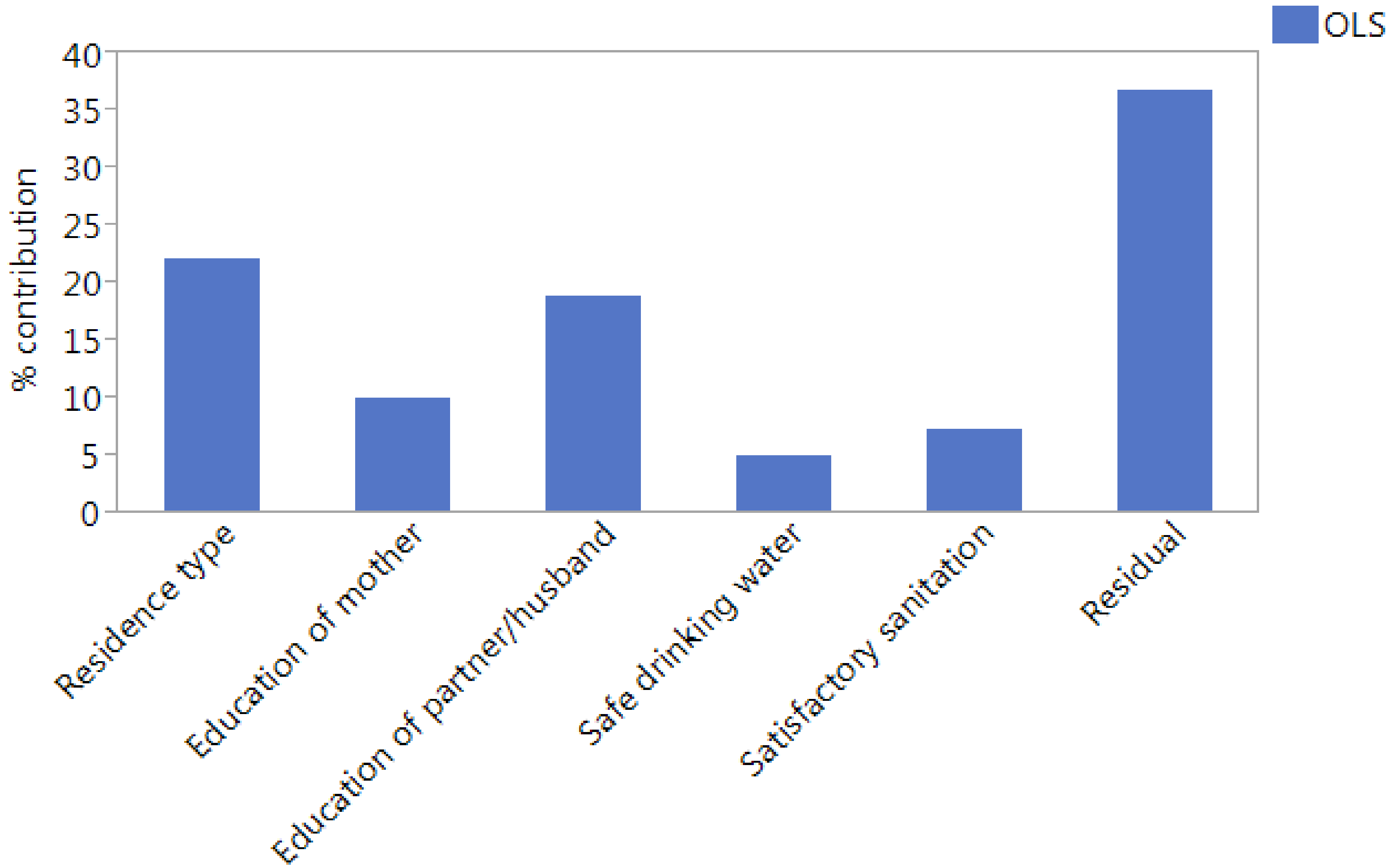
GMM vs. OLS Regression for the SEM

	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◇	-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***		–

Decomposition (I)



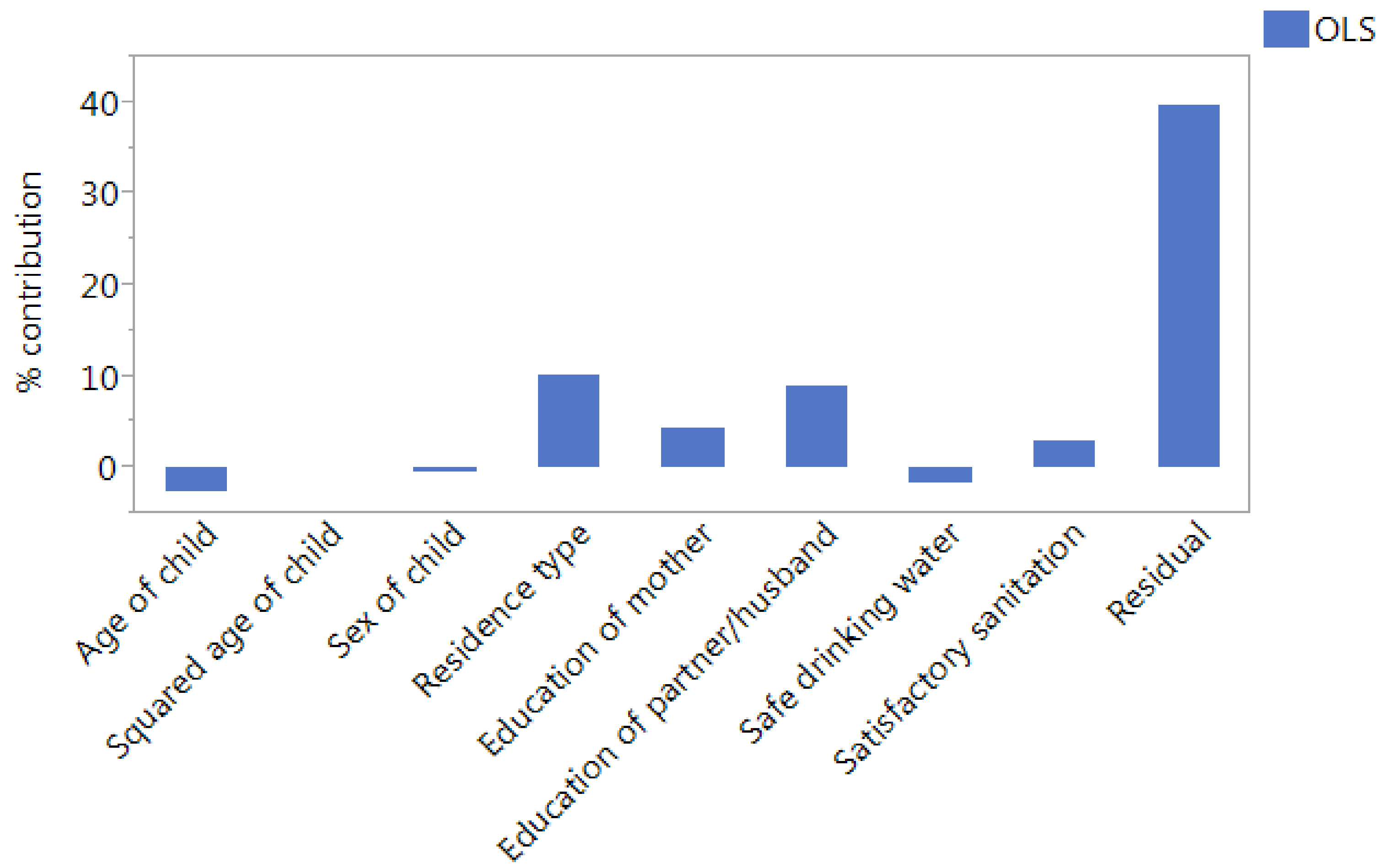
Decomposition (II)



Decomposition (III)

	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							

Decomposition (III) – Direct Effects



Results

- The GMM analysis of the SEM confirms previous findings that health is largely influenced by SES (= *d*), but the opposite relationship does not hold
  - The effect of SES on health is indirect and measured by the instruments “residence type” and “satisfactory sanitation”
- The contribution of SES (= *d*) in decomposition (I) is 42.62%, which is by far the largest
  - The contribution is indirect and measured by the variables “residence type” and “satisfactory sanitation”
  - The residual term is not zero, but equal to 38.11%

Summary

- Decomposition (III) based on the bivariate multiple regression model is also the decomposition from a SEM
- The SEM proposed is an observed-variables SEM
- Further research will involve
  - the construction of a SEM where the endogenous variables are not observed, but latent
  - indices based on socioeconomic levels rather than ranks (Erreygers & Kessels, 2014, in progress)