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
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, \ldots, \( n \))
- 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

\[ GC = 2 \text{Cov}(h, d) \]
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)

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

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

\[ d_i = \gamma_0 + \gamma_1 z_{1,i} + \gamma_2 z_{2,i} + \ldots + \gamma_q z_{q,i} + \xi_i \]

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

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

\[ GC = 2 \sum_{j=1}^{p} \lambda_j \pi_j Var(s_j) + 2 \sum_{j=1}^{p} \sum_{g=j+1}^{p} (\lambda_j \pi_g + \lambda_g \pi_j) Cov(s_j, s_g) + 2 Cov(\psi, \chi) \]
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 $\beta_1$ equal to $\frac{\text{Cov}(h,d)}{\text{Var}(d)}$ so that

\[
GC' = 2 \frac{\text{Cov}(h,d)}{\text{Var}(d)} \text{Cov}(d,d) + 2 \text{Cov}(\varepsilon, d) \\
= 2 \text{Cov}(h,d) + 0
\]
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

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

**GC = -0.0136**
## GMM vs. OLS Regression for the SEM

<table>
<thead>
<tr>
<th></th>
<th>$h$</th>
<th></th>
<th></th>
<th>$d$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>Coefficient</td>
<td>$t$-stat</td>
<td>OLS</td>
<td>Coefficient</td>
<td>$t$-stat</td>
</tr>
<tr>
<td>Constant</td>
<td>0.1187</td>
<td>13.52***</td>
<td></td>
<td>-0.1700</td>
<td>-16.01***</td>
<td></td>
</tr>
<tr>
<td>Age of child</td>
<td>0.0017</td>
<td>11.18***</td>
<td></td>
<td>-0.1493</td>
<td>-26.15***</td>
<td></td>
</tr>
<tr>
<td>Squared age of child</td>
<td>-0.0001</td>
<td>-13.48***</td>
<td></td>
<td>-0.1493</td>
<td>-26.15***</td>
<td></td>
</tr>
<tr>
<td>Sex of child</td>
<td>0.0143</td>
<td>2.41*</td>
<td></td>
<td>-0.1493</td>
<td>-26.15***</td>
<td></td>
</tr>
<tr>
<td>Residence type</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Education of mother</td>
<td>-0.0022</td>
<td>-1.81°</td>
<td>-3.36***</td>
<td>0.2457</td>
<td>21.94***</td>
<td>0.2457</td>
</tr>
<tr>
<td>Education of partner/husband</td>
<td>-0.0014</td>
<td>-1.27</td>
<td>-2.63***</td>
<td>0.0102</td>
<td>7.80***</td>
<td>0.0102</td>
</tr>
<tr>
<td>Safe drinking water</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Satisfactory sanitation</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>$d$</td>
<td>-0.0987</td>
<td>-3.46***</td>
<td>-0.0559</td>
<td>-4.67***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h$</td>
<td>—</td>
<td>—</td>
<td></td>
<td>—</td>
<td>0.0826</td>
<td>1.25</td>
</tr>
</tbody>
</table>

### Model-fit Statistics

- $R^2$: 0.0767 (GMM), 0.0796 (OLS)
- $J$: 0.42 (GMM), 2.69 (OLS)
- Cragg-Donald $F$: 917.43*** (GMM), 194.31*** (OLS)
Decomposition (I)
Decomposition (II)
### Decomposition (III)

<table>
<thead>
<tr>
<th></th>
<th>Direct effect</th>
<th>Combined effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age child</td>
<td>Squared age child</td>
</tr>
<tr>
<td>Age child</td>
<td>-2.49</td>
<td>-0.19</td>
</tr>
<tr>
<td>Squared age child</td>
<td>0.04</td>
<td>-0.19</td>
</tr>
<tr>
<td>Sex child</td>
<td>-0.32</td>
<td>-0.02</td>
</tr>
<tr>
<td>Residence type</td>
<td>10.05</td>
<td>0.15</td>
</tr>
<tr>
<td>Education mother</td>
<td>4.41</td>
<td>0.54</td>
</tr>
<tr>
<td>Education partner</td>
<td>8.99</td>
<td>0.92</td>
</tr>
<tr>
<td>Safe water</td>
<td>-1.57</td>
<td>-0.46</td>
</tr>
<tr>
<td>Satisfactory sanitation</td>
<td>3.03</td>
<td>0.21</td>
</tr>
<tr>
<td>Component total</td>
<td>22.13</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>39.76</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Decomposition (III) – Direct Effects

- Age of child
- Squared age of child
- Sex of child
- Residence type
- Education of mother
- Education of partner/husband
- Safe drinking water
- Satisfactory sanitation
- Residual

% contribution
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)