Poverty, Deprivation, Exclusion:
A Structural Equations Modelling Approach

Ottó Hajdu

Head of Department
Corvinus University
of Budapest

E-mail: hajduotto@uni-corvinus.hu

The main purpose of the paper is to build a path model of latent variables: poverty, relative deprivation and social exclusion. Applying the SEM methodology, the poverty is measured in a multivariate approach eliminating the need to cut off the poor's subset by a fixed poverty line. The focus is on estimating hypothetical structural path coefficients on the one hand and testing their significance on the other. The computations are carried out using the Statistica 8.0 software based on the 2003 Hungarian Household Survey, taking the household as the unit of observation.

KEYWORDS:
SEM models.
Poverty.
Deprivation.

In our conception, the constructions of poverty, relative deprivation and social exclusion in a society constitute a multivariate structural equations system of latent variables. The endogenous variables in this hypothetical latent causality structure can be considered also exogenous in other equations of the path diagram. (See Figure 1.) The paper suggests a way to estimate and test hypothetical relations among these constructions with no need to split the society into "the poor" and "the non-poor" clusters by a strict poverty line. Obviously, the meaning of the latent constructions is given by their measured manifest variables, which can also be distinguished either endogenous or exogenous.

1. Conception

The initial conceptual model is described in Figure 1. Considering the path diagram, boxes identify manifest variables, while ovals show latent variables.

Figure 1. Initial multiple indicator multiple cause model
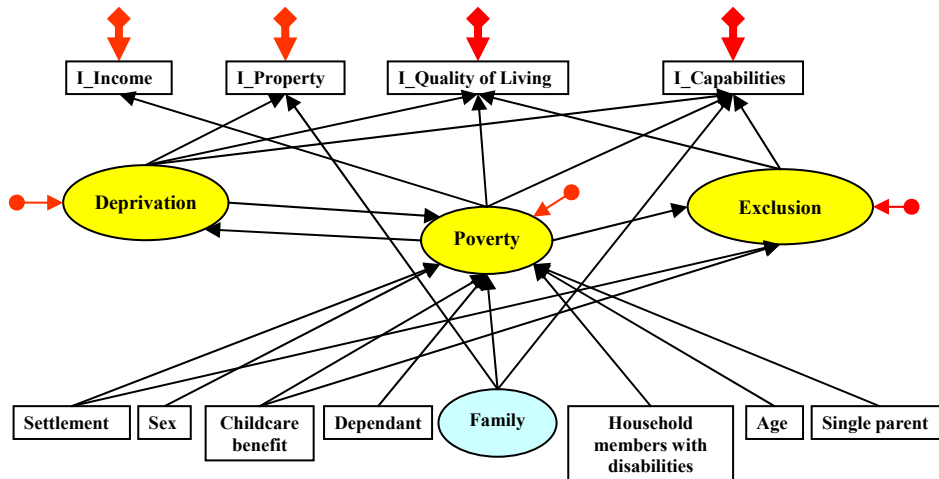
Note. The arrow ●→ stands for a residual latent variable pointing to its own latent variable, and it is termed EPSILON# in Table 1. The arrow ←→ marks a set of residual latent variables corresponding to their own manifest variables but they are not included in Table 1. Manifest variables without a residual variable are obviously exogenous.

An arrow between two variables represents a regression coefficient. The main goal is to estimate and then test these parameters. Especially, a manifest variable explained by a latent variable is termed "indicator". An "I_" type box with an arrow pointing to it indicates a set of indicators corresponding to the latent variable connected with. Some indicators are allowed to belong to several I_ boxes. Furthermore, each endogenous variable has an error (residual) variable represented by a single arrow pointing to it.

Considering the *latent part* of the model, we define the following constructions. Firstly, the endogenous variables are:

1. Poverty: the household is living in poverty (for example at the bottom of the income scale);

2. Deprivation: the household is deprived of several goods therefore its members may feel poor compared to the richer ones;

3. Exclusion: the household is excluded from certain socio-economic functioning.

The only exogenous latent variable is:

4. Family: the family background behind the household (for example the extent to which the household is supported by the whole family).

Secondly, regarding the manifest variables, the exogenous variables (with their scale in a square bracket) are:

1. Settlement [capital, large city, city, village]: the type (level) of the settlement where the household lives;

2. Sex [male, female]: the gender of the head of the household;

3. Childcare benefit [yes, no]: whether the head of the household receives a childcare benefit (allowance);

4. Dependant(s) [0, 1, 2 or more]: the number of dependants younger than 25 in the household;

5. Household member(s) with disabilities [yes, no]: whether there is a permanently sick person in the household;

6. Age [20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75]: the age class of the head of the household as defined by the lower bounds;

7. Single parent [yes, no]: single parent with a child or children.

Finally, the considered endogenous manifest indicators (grouped by their contents) are as follows (as mentioned earlier, "I_" stands for "Indicator Set"):

8. I_Income variables:
   a) Income per capita: annual per capita income (thousand HUF);
9. I_Property variables:
   a) Flat_HUF: the value of the flat assessed by the head (million HUF);
   b) Flat2 [yes, no]: whether a member of the household owns a second flat;
10. I_Quality of living variables:
    a) Running water [yes, no]: whether there is running water in the household;
    b) Flat problems: the number of several types of defects experienced in the flat;
    c) Durables: the number of durables in the household;
    d) Environment: qualities of the environment surrounding the flat;
    e) Car_HUF: the value of the car of the household assessed by the head (million HUF);
    f) Car_Km: the annual performance of the car of the household (thousand kilometres);
    g) Public health: the number of household members eligible for public health service treatment/benefits;
11. I_Capability variables:
    a) Education [1, 2,…,13]: the educational attainment of the head of the household (for example 1 means uncompleted elementary school, while 13 is for PhD degree);
    b) Unemployed household member(s) [0, 1, 2, 3 or more]: the number of unemployed persons in the household;
    c) Default: The number of the types of earlier unpaid bills;
    d) Retired household member(s): The number of pensioners in the household.

Based on the data of 3 571 Hungarian households of 2003, the corresponding parameter estimation results (coefficients, standard errors, *t*-statistics and their probability significance values) are included in Table 1. According to these results, the *insignificant relations* – at 10 percent significance level – are as follows.

Table 1

Parameter estimation results

Parameter	Coefficient	Standard Error	t	Probability
Structural part of the model				
(POVERTY)-1->(DEPRIVATION)	−15.441	0.034	−452.015	0.000
(POVERTY)-2->(EXCLUSION)	−0.866	0.036	−23.870	0.000
(DEPRIVATION)-3->(POVERTY)	1.493	0.073	20.442	0.000
(FAMILY)-4->(POVERTY)	−0.071	0.044	−1.622	*0.105*
(EPSILON1)-5-(EPSILON1)	1.244	0.213	5.847	0.000
(EPSILON2)-6-(EPSILON2)	258.943	0.000		
(EPSILON3)-7-(EPSILON3)	0.249	0.062	3.995	0.000
Exogenous manifest part of the model				
[Settlement]-8->(POVERTY)	−0.331	0.040	−8.290	0.000
[Settlement]-9->(EXCLUSION)	0.034	0.015	2.360	0.018
[Sex]-10->(POVERTY)	0.098	0.068	1.451	*0.147*
[Childcare benefit]-11->(POVERTY)	0.013	0.070	0.182	*0.855*
[Childcare benefit]-12->(EXCLUSION)	0.026	0.029	0.910	*0.363*
[Dependants]-13->(POVERTY)	0.107	0.026	4.033	0.000
[Household member(s) with disabilities]-14->(POVERTY)	0.005	0.061	0.079	*0.937*
[Age of the head of the household]-15->(POVERTY)	−0.018	0.011	−1.624	*0.104*
[Single parent]-16->(POVERTY)	0.297	0.131	2.261	0.024
Indicator manifest part of the model				
(POVERTY)-17->[Running water]	−0.079	0.001	−76.911	0.000
(POVERTY)-18->[Default]	0.022	0.012	1.888	0.059
(POVERTY)-19->[Flat problems]	−2.088	0.152	−13.778	0.000
(POVERTY)-20->[Unemployed household member(s)]	−0.642	0.093	−6.890	0.000
(POVERTY)-21->[Retired household member(s)]	−0.012	0.010	−1.185	*0.236*
(POVERTY)-22->[Income per capita]	40.312	6.842	5.892	0.000
(DEPRIVATION)-23->[Durables]	0.237	0.071	3.337	0.001
(DEPRIVATION)-24->[Car_HUF]	−339.757	25.721	−13.209	0.000
(DEPRIVATION)-25->[Flat_HUF]	1.430	0.139	10.293	0.000
(DEPRIVATION)-26->[Flat2]	−0.086	0.016	−5.496	0.000
(DEPRIVATION)-27->[Environment]	0.273	0.021	13.147	0.000
(DEPRIVATION)-28->[Unemployed household member(s)]	0.208	0.023	8.898	0.000
(DEPRIVATION)-29->[Education]	−2.830	0.141	−20.094	0.000
(EXCLUSION)-30->[Car_Km]	−0.286	0.182	−1.575	*0.115*
(EXCLUSION)-31->[Public Health]	−0.006	0.004	−1.590	*0.112*
(EXCLUSION)-32->[Unemployed household member(s)]	−0.592	0.089	−6.623	0.000
(EXCLUSION)-33->[Education]	−2.167	0.148	−14.658	0.000
(FAMILY)-34->[Flat2]	0.615	0.008	80.979	0.000
(FAMILY)-35->[Education]	-0.140	0.077	−1.805	0.071

Note: The "-#-" wire between two EPSILON# variables indicates covariance, or variance when computed as a self-covariance.

In the structural latent part:

> *1*. "Poverty" regressed on "Family background",

In the exogenous manifest part:

> *2*. "Poverty" regressed on "Sex";
> *3*. "Poverty" regressed on "Childcare benefit";
> *4*. "Poverty" regressed on "Household member(s) with disabilities";
> *5*. "Poverty" regressed on "Age of the head of the household";
> *6*. "Exclusion" regressed on "Childcare benefit";

In the indicator manifest part:

> *1*. "Retired household member(s)" regressed on "Poverty";
> *2*. "Car_Km" regressed on "Exclusion";
> *3*. "Public Health" regressed on "Exclusion".

Based on the model, the following conclusions were drawn from the structural latent part: *1*. if someone is poor, then he/she is consequently deprived and vice versa; *2*. if someone is poor, then he/she accordingly excluded; *3*. the family background does not have impact on the poverty status of the household;

from the exogenous manifest part: *4*. the poverty level is not influenced by the sex and age of the head of the household or childcare benefit assistance or the existence of a sick person in the household; *5*. the settlement type, the number of dependants younger than 25 years in the household and the single-parent household structure affect the poverty level; *6*. unlike childcare benefit assistance, the type of the settlement determines the level of exclusion;

from the indicator manifest part: *7*. retirement is not an indicator of the poverty level; *8*. annual performance and *9*. eligibility for public health service do not provide an indication of the exclusion level.

2. A methodological overview
of the asymptotically distribution free estimator

Due to the fact that the socio-economic measurement variables are rarely distributed normally, a brief overview of the so-called asymptotically distribution free estimator methodology of Structural Equation Models (SEMs) applied previously is given as follows.

2.1. The structural equations path model

Collecting all the variables considered in the vector **v**, the structural equations path model (SEPATH) takes the form

$$\mathbf{v} = \begin{bmatrix} \mathbf{y}_L \\ \mathbf{y}_M \\ \mathbf{x}_M \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{LL} & \mathbf{B}_{LM} & \mathbf{0} \\ \mathbf{B}_{ML} & \mathbf{B}_{MM} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \times \begin{bmatrix} \mathbf{y}_L \\ \mathbf{y}_M \\ \mathbf{x}_M \end{bmatrix} + \begin{bmatrix} \mathbf{G}_{LM} & \mathbf{G}_{LL} \\ \mathbf{G}_{MM} & \mathbf{G}_{ML} \\ \mathbf{I}_{MM} & \mathbf{0} \end{bmatrix} \times \begin{bmatrix} \mathbf{x}_M \\ \mathbf{x}_L \end{bmatrix} \qquad /1/$$

or briefly we get

$$\mathbf{v} = \mathbf{Bv} + \mathbf{Gx} \qquad /2/$$

from which the reduced form is as follows:

$$\mathbf{v} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{Gx}, \qquad /3/$$

where **y** and **x** are vectors of endogenous and exogenous variables respectively, L and M subscripts indicate latent and manifest variables respectively, **I** is the identity matrix, and finally, **B** and **G** are corresponding matrices of structural coefficients to be estimated. The number of manifest variables is p.

Because the estimation procedure is based on fitting the distinct sample covariances among the manifest variables as a function of the parameters to be estimated, extraction of the manifest variables included in the vector $\mathbf{m}_p$ – as a function of the exogenous variables only – is necessary. Firstly, the filtering step is

$$\mathbf{m}_p = \underbrace{\begin{bmatrix} \mathbf{0}_{p,l} & \mathbf{I}_{p,p} \end{bmatrix}}_{\mathbf{D}} \times \begin{bmatrix} \mathbf{l}_l \\ \mathbf{m}_p \end{bmatrix} = \mathbf{Dv}, \qquad /4/$$

where **D** is a dummy-type filter matrix, and finally, with the substitution of **v**,

$$\mathbf{m} = \mathbf{D}(\mathbf{I} - \mathbf{B})^{-1} \mathbf{Gx}. \qquad /5/$$

2.2. Parameter estimation and identification

Now, using some covariance algebra, the covariance matrix $\mathbf{C}_{m,m}$ of the manifest variables of order (p,p) can be expressed as a function of the coefficient matrices and the covariance matrix $\mathbf{C}_{xx}$ of the exogenous variables:

$$\mathbf{C}_{m,m} = \left(\mathbf{D}(\mathbf{I}-\mathbf{B})^{-1}\mathbf{G}\right)\mathbf{C}_{xx}\left(\mathbf{G}^T(\mathbf{I}-\mathbf{B})^{-1^T}\mathbf{D}^T\right),\qquad /6/$$

where T means transposed.

A single covariance between variables j and t can now be derived from the model by a nonlinear f_{jt} function as follows:

$$Cov\left(m_j,m_t\right)=f_{jt}\left(\mathbf{B},\mathbf{G},\mathbf{Cov}_{xx}\right),\quad (j,t)=1,2,...,p\left(p+1\right)/2\qquad /7/$$

or

$$Cov\left(m_j,m_t\right)=f_{jt}\left(\theta_1,\theta_2,...,\theta_q\right),\qquad /8/$$

where the vector $\boldsymbol{\theta}=\left(\theta_1,\theta_2,...,\theta_q\right)$ contains the free parameters to be estimated based on nonlinear equations $p\left(p+1\right)/2$ in number. Hence the degree of freedom of the SEPATH model is

$$df=\frac{p\left(p+1\right)}{2}-q-r,\qquad /9/$$

where r is the number of the endogenous latent variables.

2.3. Asymptotically distribution free estimator

The so-called quadratic form fitting function to be minimized with respect to $\boldsymbol{\theta}$ to yield $\hat{\boldsymbol{\theta}}$ is

$$F=\left(\mathbf{s}-\sigma(\boldsymbol{\theta})\right)^T\mathbf{W}\left(\mathbf{s}-\sigma(\boldsymbol{\theta})\right)\rightarrow\min,\qquad /10/$$

where $\mathbf{s}$ is a vector of order $p(p+1)/2$ consisting of the distinct (non-duplicated) elements of the sample covariance matrix $\mathbf{S}$ of order (p,p), $\sigma(\boldsymbol{\theta})$ is the corresponding same order vector of the hypothetical covariances based on the model, and finally $\mathbf{W}$ is a positive definite weight matrix of order $(p(p+1)/2, p(p+1)/2)$.

The optimal weight matrix – based on sample size N – is the sampling-distribution inverse covariance matrix of $\mathbf{s}$ denoted by $(\mathbf{C}_{ss})^{-1}$ where the covariance between two sample covariances follows from

$$\frac{1}{N-1}Cov\left(s_{jk},s_{lt}\right)=\sigma_{jl}\sigma_{kt}+\sigma_{jt}\sigma_{kl}+\frac{N-1}{N}\kappa_{jklt}\qquad /11/$$

with the kurtosis measure

$$\kappa_{jklt} = \sigma_{jklt} - \left(\sigma_{jk}\sigma_{lt} + \sigma_{jl}\sigma_{kt} + \sigma_{jt}\sigma_{kl}\right) \qquad /12/$$

based on the forth-order product moment defined as

$$\sigma_{jklt} = E\left(x_j - \mu_j\right)\left(x_k - \mu_k\right)\left(x_l - \mu_l\right)\left(x_t - \mu_t\right), \qquad /13/$$

where μ_j stands for the population mean of the variable x_j, E means expected value and j, k, l, t are for indices of manifest variables.

Apparently, in the large sample case, the elements of the asymptotic weight matrix are given as

$$w_{jk,lt} = \sigma_{jklt} - \sigma_{jk}\sigma_{lt}. \qquad /14/$$

Specially, assuming that all the variables have a common $\kappa = \sigma_{jjjj} / 3\sigma_{jj}^2 - 1$ degree of kurtosis, we obtain the so-called elliptical estimator with forth-order moment of

$$\sigma_{jklt} = \left(\kappa + 1\right)\left(\sigma_{jk}\sigma_{lt} + \sigma_{jl}\sigma_{kt} + \sigma_{jt}\sigma_{kl}\right). \qquad /15/$$

Finally, writing the fitting function in the alternative form of

$$F = \frac{1}{2}tr\left\{\left(\left[\mathbf{S} - \hat{\mathbf{S}}\right]\mathbf{V}^{-1}\right)^2\right\} \rightarrow \min, \qquad /16/$$

(*Bollen* [1989]), where *tr* is for trace, the so-called generalized least squares (GLS) estimator is used when the positive definite weight matrix $\mathbf{V}$ of order (p,p) is the sample covariance matrix $\mathbf{S}$ itself, and the iteratively reweighted GLS is applied when the weight matrix is the fitted covariance matrix $\hat{\mathbf{S}} = \mathbf{\Sigma}\left(\hat{\mathbf{\theta}}\right)$ updated in each successive iterative step based on the latest parameter estimation $\hat{\mathbf{\theta}}$.

2.4. Evaluating goodness-of-fit: Model selection

The model evaluating process means testing the distance between

1. the target (currently estimated) model and the so-called null model by the independence test (where the null model is defined with no latent variables at all);

2. the target (currently estimated) model and the so-called saturated model with a perfect fit by the goodness-of-fit test; and

3. the currently estimated target model and another candidate target model with more or less parameters by the nested models test.

Figure 2. Goodness-of-fit of nested models
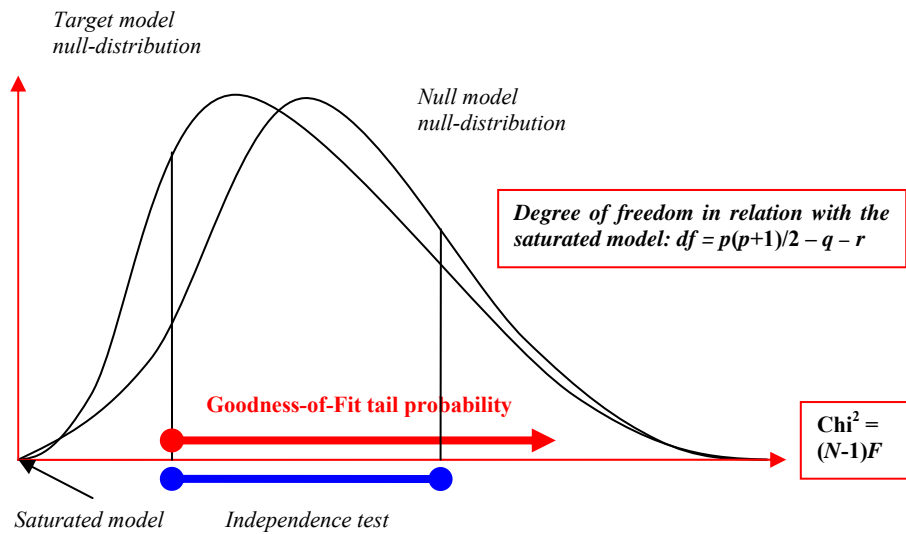
Note. The null distribution assumes that the null hypothesis is true.

The details characterizing the initial target model are:

1. The sample size: $N = 3571$;

2. The number of manifest variables: $p = 21$, the number of endogenous latent variables: $r = 3$;

3. The number of freely estimated parameters: $q = 56$;

4. The goodness-of-fit Chi-square value of the null model:
Chi2 = 37577.82 with df = 20*21 / 2 = 210;

5. The converged value of the objective function: $F = 5$;

6. The goodness-of-fit Chi-square value of the currently estimated target model: GF_Chi2 = 17850 with *df* = 231–56–3 = 172 and tail probability = 0.000.

7. The goodness-of-fit heuristic measures are presented in Table 2.

(For the meaning of these measures see *Bollen* [1989] or *Hajdu* [2003].)

Based on the small (0.000) tail probability, on the one hand, the distance from the saturated model is significant (due to the large sample size) hence indicating a poor fit. On the other hand, the values of the heuristic measures included in Table 2 show a considerable initial goodness-of-fit to be improved by a refined more complex model. (For instance, the Bentler comparative fit index measures a 52.69 percentage distance from the null model relative to the distance between the null and saturated (i.e. the two extreme) models.)

Table 2

Heuristic goodness-of-fit indices

| Index name	Index formula
Population non-centrality index *	$NCI = \dfrac{\chi^2 - df}{N-1} = 4.9504$
Steiger–Lind root mean square error *	$RMSE = \sqrt{\dfrac{1}{df}\max\{NCI,0\}} = 0.1697$
McDonald non-centrality index	$MDNI = \exp\{-0.5\max(NCI,0)\} = 0.0841$
Population gamma index	$\Gamma_1 = \dfrac{p}{2NCI+p} = 0.6796$
Adjusted population gamma index	$\Gamma_2 = 1 - \dfrac{p(p+1)}{2df}(1-\Gamma_1) = 0.3621$
Jöreskog–Sörbom GFI	$GFI = 1 - \dfrac{2F}{tr\left(\left[\mathbf{S}\hat{\Sigma}^{-1}\right]^2\right)} = 0.5240$
Adjusted Jöreskog–Sörbom	$AGFI = 1 - \dfrac{p(p+1)}{2df}(1-GFI) = 0.3621$
Akaike information criterion *	$AC = F + \dfrac{2q}{N-1} = 5.0314$
Schwarz's Bayesian criterion *	$SC = F + \dfrac{q\ln(N)}{N-1} = 5.1283$
Browne–Cudeck cross validation index *	$CV = F + \dfrac{2q}{N-p-2} = 5.0316$
Bentler–Bonett, Tucker–Lewis non-normed fit index	$NNFI_{t/b} = 1 - \dfrac{df_b}{df_t}\dfrac{\chi_t^2 - df_t}{\chi_b^2 - df_b} = 0.4224$
Bentler comparative fit index	$BCFI_{t/b} = 1 - \dfrac{\chi_t^2 - df_t}{\chi_b^2 - df_b} = 0.5269$
Bollen's Rho	$\rho_{t/b} = 1 - \dfrac{df_b}{df_t}\dfrac{\chi_t^2}{\chi_b^2} = 0.4200$

* The indices indicated by an asterisk select the preferred model at their minimized values.

Note. Sample size = N; p = the number of manifest variables; q = the number of free parameters. Subscription t indicates the target (more complex) model and b stands for the baseline null model; $F = \chi^2/(N-1)$ is the converged value of the "fitting function".

3. An extended model

A hypothetical extended and modified candidate model is suggested in Figure 3.

Figure 3. The extended hypothetical structural model
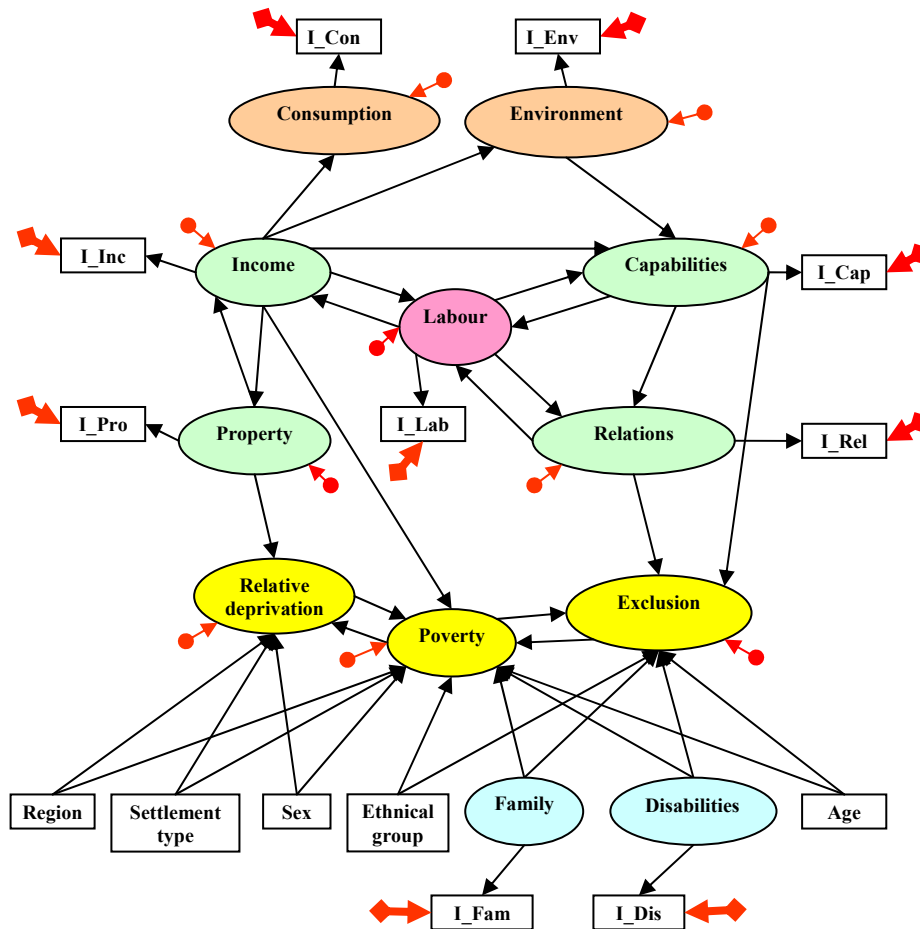
The list of the latent variables is as follows. Firstly, the endogenous variables defined are: *1*. Poverty; *2*. Relative deprivation; *3*. Social exclusion; *4*. Labour; *5*. Property; *6*. Income; *7*. Consumption; *8*. Environment; *9*. Capabilities; *10*. Relations.

Secondly, the exogenous variables are: *11*. Family; *12*. Disabilities.

Thirdly, the exogenous manifest variables considered are: *13*. Region; *14*. Settlement type; *15*. Sex of the head of the household; *16*. Ethnical group; *17*. Age group of the head of the household.

Finally, the endogenous manifest indicators (all of them can consist of several items, and note that "Poverty", "Deprivation" and "Exclusion" have no direct indicators) are: *18*. Indicators of income level (I_Inc); *19*. Indicators of properties (I_Pro); *20*. Indicators of capabilities (I_Cap); *21*. Indicators of relations (I_Rel); *22*. Indicators of labour (I_Lab); *23*. Indicators of consumption (I_Con); *24*. Indicators of the environment (I_Env); *25*. Indicators of the family background (I_Fam); *26*. Indicators of being disabled (I_Dis).

The hypothesis behind this model – among other candidates – can be computed and tested based on the model selection methodology described previously. These computations are not presented in this paper, they shall be subject of further investigation.

4. Conclusions

The paper suggests a new way of testing relationships among the constructions of poverty, deprivation and social exclusion with no need to split the society into "the poor" and "the non-poor". Hence, these latent dimensions can be measured based on several socio-economic variables using a multivariate approach. The paper gives an initial latent structure model and then suggests a refined version of it. In addition, in order to select among competing models, a brief overview of goodness-of-fit methodology of SEMs with the parameter estimation method behind it is also presented.

References

BOLLEN, K. A. [1989]: *Structural Equations with Latent Variables*. Wiley. New York.

ÉLTETŐ, Ö. – HAVASI, É. [2006]: Poverty in Hungary with Special Reference to Child Poverty. *Hungarian Statistical Review*. Vol. 84. Special No. 10. pp. 3–17.

HAJDU, O. [2003]: A kovariancia-struktúra modellek illeszkedésvizsgálata. *Statisztikai Szemle*. Vol. 81. No. 5–6. pp. 442–465.

MONOSTORI, J. [2002]: Social Relationships of the Poor. *Hungarian Statistical Review*. Vol. 80. Special No. 7. pp. 18–44.