A Method for The Estimation of The Distribution of Human Capital from Sample Surveys on Income and Wealth

Giorgio Vittadini, University of Bicocca Milan, Italy **Camilo Dagum**, University of Bologna, Italy and University of Ottawa, Canada, **Pietro Giorgio Lovaglio**, University of Bicocca-Milan, Italy, **Michele Costa**, University of Bologna, Italy.
Giorgio Vittadini, Department of Statistics, University of Bicocca-Milan,Via Bicocca degli Arcimboldi 8, 20126 MILAN, ITALY.

KEY WORDS: Human Capital, Latent Variable, Pls, Lisrel, Optimal Scaling

Introduction

This paper proposes a method for estimating the 1983 U.S. household distribution of Human Capital. From the statistical point of view, the HC is defined as a Latent Variable measured by a set of observed mixed indicators in a Path Analysis Model. The HC estimates consider the definitions advanced for a Latent Variable in a Path Analysis with respect to formative and reflective indicators.

The set of indicators and their links with HC

The concept of Human Capital (HC), theoretically and systematically developed over the last 50 years (Mincer 1958, 1970; Becker 1962, 1964 and Schultz 1959, 1961) has been estimated in literature by either the retrospective (Kendrick 1976; Eisner, 1985) or prospective methods (Jorgenson and Fraumeni 1989). The first, dealing with the cost of production, is insufficient for various reasons, because it does not take into account the social costs, such as public investment in education, the variables concerning home conditions and community environments, and the genetic contribution to HC, including health conditions (Dagum and Vittadini 1996). Moreover, the actual effects of the investment in HC on the income and wealth of the households are not considered.

In the prospective method the HC can be defined as the present actuarial value of an individual's expected income related to his skill, acquired abilities, and education (Dagum and Slottje 2000). However, the prospective method reduces the **HC** investment to its monetary value in terms of an assumed flow of income, and it ignores the amount of investment in education, job training and other investments. It is also difficult to predict future income.

We now present a new methodology to estimate the distribution of HC in families, giving greater emphasis to economic issues because the definition of HC involves both its investment amounts on families and its effect on income. In this case, instead of quantitative financial indicators, we have a composite set of qualitative and quantitative indicators (Table 1) with the Path Analysis diagram showing their causal links.

The "indirect" set of indicators $\Gamma = (\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_6, \mathbf{y}_8, \mathbf{y}_9, \mathbf{y}_{10}, \mathbf{y}_{11}, \mathbf{y}_{12}, \mathbf{y}_{13})$ involved in the Path Analysis is composed of a set of causal links between themselves and a set of indicators $[\Psi = (\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7, \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7, \mathbf{y}_{15})]$ and $\mathbf{y}_{14}$ (total wealth), which measure the investment in education and are directly connected with **HC** in (2).

As proposed in statistical literature (Tenenhaus, 1995), these indicators can be defined as formative indicators because they "form" or "cause" the multidimensional construct HC in equation (2). **HC** is a "consequence" of the investment in education. It is measured by a set of formative indicators $\mathbf{F} = (\mathbf{y}_{14}, \Psi)$

The most important formative indicators are years of education, years of full time and part time employment, and wealth. Marital status, gender, region and age are involved as well, because the actual value of investment in HC is influenced by these personal and environmental conditions.

The effects on income of the households HC and wealth are presented in (3). Therefore $\mathbf{y}_{17}$ can be classified as reflective indicator, because it reflects HC, in the sense that it is a consequence of the amount and type of the investment in education. Wealth ($\mathbf{y}_{14}$) is both a formative indicator and an independent cause of income.

The econometric specification and analysis was made by Dagum (1994) and Dagum et al. (2003).

Indirect Indicators Γ	
x_2=H Gender; x_3 = H Race; x_6 = S Age; y_5 = H Years of Not Full-Time Work; y_7 = S Years of Not Full-Time Work; y_8 = H Job Status; y_9 = H Occupation; y_{10}= H Industry; y_{11}= S Job Status; y_{12}= S Occupation y_{13} = S Industry	
Formative indicators F = (Ψ, y_{14})	
Ψ: x_1 = H Age; x_4= Region ;x_5 = H Marital Status; x_7= S Gender; y_1= H Years of Schooling; y_2 = S Years of Schooling; y_3 = Number of Children; y_4 = H Years of Full-Time Work; y_6= S Years of Full-Time Work; y_{14}= Household Total Wealth; y_{15} = Household Total Debts.	
Reflective indicator y_{17} = Household Income	**H:** Household Head; **S:** Spouse

Table 1 Observed indicators

The statistical definition of the LV HC

We have already stated (Dagum and Vittadini 1996) that, from the statistical point of view, HC can be expressed as an LV. But there are different ways an LV can be defined. Traditionally, a variable can be defined as an LV if the equations cannot be manipulated into expressing the variable as a function of manifest variables (Bentler 1982). In other words, in this definition, an LV is a factor that underlies and causes reflective indicators and accounts for their observed variance in a measurement model (typically the factor model) given the effects of other explicative indicators (in this case the reflective indicator Income, given the effect of the explicative indicator wealth in equation (3)). Otherwise we can define HC as a latent variable caused and measured (with errors) by a linear combination of the formative indicators **F** in equation (2). Finally we can propose a third, more complete, definition of an LV, as in this case where it is connected with both formative and reflective indicators in a Path Diagram. Hence the latent variable HC can be defined as a linear combination of formative indicators **F** that best fits the reflective indicator earning income, as in equations (2)-(3).

The proposed methodology

This approach completes the methodology proposed by Dagum and Slottje (2000) where they combine a zerodimensional latent variable approach (part A) and an actuarial mathematical approach (part B).

The Latent Variable approach proposes a new methodology able to obtain the zerodimensional HC latent variable, then transforms the estimated latent variable into an accounting monetary value, and finally estimates the mean value of **HC**. The Path Analysis and the Latent Variable Approach are shown in Figure1.

The Actuarial Mathematical approach starts with the actuarial estimation, in monetary values, of the average human capital by age of economic units and finally estimates the average of the population in monetary units. The synthesis gives the final HC estimation and distribution of American Household.

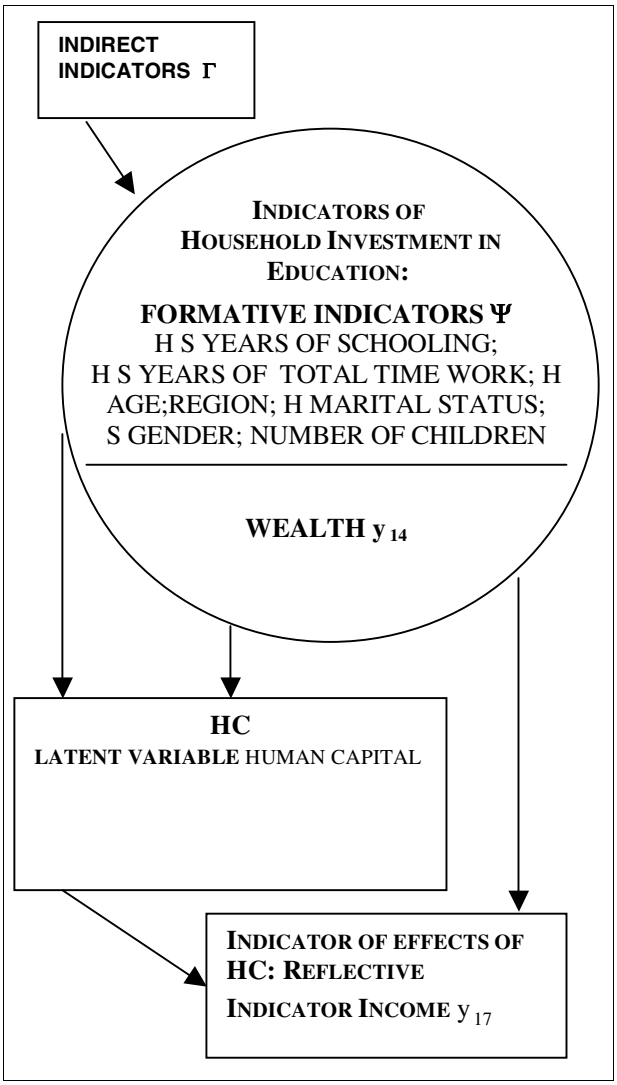

Figure 1: Path Analysis and Latent Variables approach

$$
\begin{aligned}
y_1 &= g_1(x_1, x_3, x_4, x_5) + u_1 \\
y_2 &= g_2(x_1, x_2, x_3, x_4, x_5, y_1) + u_2 \\
y_3 &= g_3(x_1, x_3, x_4, x_5, y_1) + u_3 \\
y_4 &= g_4(x_1, x_2, x_3, x_4, x_5, y_2, y_3) + u_4 \\
y_5 &= g_5(x_1, x_2, x_3, x_4, x_5, y_1, y_4) + u_5 \\
y_6 &= g_6(x_2, x_3, x_4, x_5, x_6, y_2, y_3, y_4) + u_6 \\
y_7 &= g_7(x_2, x_4, x_5, x_6, y_2, y_5, y_6) + u_7 \\
y_8 &= g_8(x_1, x_2, x_3, x_4, x_5, y_1, y_3, y_4) + u_8 \qquad (1) \\
y_9 &= g_9(x_1, x_2, x_3, y_8) + u_9 \\
y_{10} &= g_{10}(x_2, x_3, x_4, y_1, y_4, y_5, y_9) + u_{10} \\
y_{11} &= g_{11}(x_1, x_2, x_4, x_5, x_6, y_2, y_3, y_6, y_9) + u_{11} \\
y_{12} &= g_{12}(x_1, x_2, x_4, x_5, x_6, y_2, y_3, y_9, y_{11}) + u_{12} \\
y_{13} &= g_{13}(x_3, x_4, y_6, y_{12}) + u_{13} \\
y_{14} &= g_{14}(x_4, y_1, y_2, y_4, y_7, y_8, y_9, y_{10}, y_{11}, y_{12}, y_{13}) + u_{14} \\
y_{15} &= g_{15}(x_1, x_3, x_4, y_1, y_2, y_3, y_4, y_9, y_{10}, y_{12}, y_{14}) + u_{15}
\end{aligned}
$$

$$HC = Fg = [y_{14}, \Psi]\, g + u_{16} \qquad (2)$$

$$\mathbf{y_{17}} = \mathbf{y_{14}}\, k_1 + \mathbf{HC}\, k_2 + \mathbf{u_{17}} \qquad (3)$$

The Latent Variable Approach: previous proposal

The traditional proposal in statistical literature is to obtain the latent variable **HC** as a latent cause which underlies observed indicators by means of Factor Analysis . In this case starting from (3), we obtain:

$$\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{y_{17}} = \mathbf{HC}\, k_2^{\#} + \mathbf{u_{17}} \qquad (4)$$

where $\mathbf{Q}_{\mathbf{y}_{14}} = \mathbf{I} - \mathbf{P}_{\mathbf{y}_{14}}$ is the orthogonal complement of the column spaces of $\mathbf{y_{14}}$ with $\mathbf{P}_{\mathbf{y}_{14}} = \mathbf{y_{14}}(\mathbf{y_{14}}'\,\mathbf{y_{14}})^{-1}\,\mathbf{y_{14}}'$.

By means of Factor Analysis we obtain **HC** as the latent cause of the reflective indicator earning income**.** First of all, in this way we define the **HC** without taking into account the amount of investment in education measured by the formative indicators **F.** Secondly, under general conditions, given earning income Wealth $\mathbf{Q}_{\mathbf{y}_{14}}\,\mathbf{y_{17}}$, the parameter $k_2^{\#}$ is not identified and the scores of the latent variable **HC** are not unique**.** In a Factorial or in a Structural model when the expected values of latent variables are null, the identification problem is essentially whether or not vector ϑ of parameters and of variances and covariances of latent variables and errors is uniquely determined by the covariance matrix $\mathbf{\Sigma}$ of indicators whose elements are σ_{ij}. In other words if a vector ϑ can be uniquely determined from $\mathbf{\Sigma}$ (and therefore if $\mathbf{\Sigma}$ is generated by one and only one vector ϑ) then solving the equations $\sigma_{ij} = \sigma_{ij}(\vartheta)$, $i \le j$ (with p manifest variables, there are ½ p(p + 1) equations in n(θ) unknown parameters), or a subset of them, this vector of parameter is identified and the whole model is said to be identified; otherwise it is not. Anderson and Rubin showed that a necessary condition for identification is that the number of equations $\sigma_{ij} = \sigma_{ij}(\vartheta)$, $i \le j$ must be greater than the order of the vector ϑ: $p \ge 2t\ n(\theta)+1$. However, since the equations above are often non-linear, the solution is often complicated and tedious, and explicit solutions for all ϑ's seldom exist. "No general and practically useful necessary and sufficient conditions for identification are available" (Everitt 1984).

If a model is not completely identified, appropriate restrictions may be imposed on ϑ to make it identifiable. The choice of restrictions may affect the interpretation of the results of an estimated model. Under general conditions for the Factor Model, if we do not consider a few very restricted cases in which conditions for identifiability are studied analytically, e.g. where the endogenous variables are measured without error (Geraci 1976), the problem cannot be resolved. In practice, it is suggested (Jöreskog, 1981b) that "The identification problem can be studied on a case by case basis by examining the equations", choosing the restriction, not only in number but also in position, in order to obtain unique solutions. This is also true in the case of local identifiability of the parameters (Wegge 1965, 1991 Fisher 1976, Rothenberg 1971, Geraci 1976, Bekker and Pollock 1986, Shapiro 1985, Bekker 1989, 1991, Wegge and Feldman, 1983).

In our case, we have one equation $\sigma_{ij} = \sigma_{ij}(\vartheta)$:

$$\boldsymbol{\sigma}_{\mathbf{Q}_{\mathbf{y}_{14}}\,\mathbf{y}_{17}} = (k_2^{\#})^2 + \sigma_{u17} \qquad (5)$$

With two unknown values, the square of the parameter $k_2^{\#}$ $(k_2^{\#})^2$ and the variance of the error $\mathbf{u_{17}}$ (σ_{u17}). Therefore, under general conditions, when the Reliability Ratio between $\boldsymbol{\sigma}_{\mathbf{Q}_{\mathbf{y}_{14}}\,\mathbf{y}_{17}}$ and $(k_2^{\#})^2$ is unknown or the variance of the error $\sigma_{\mathbf{u17}}$ or Instrumental Variables are not available, the model (4) is not identifiable (Fuller, 1987).

Regarding the problem of indeterminacy we can verify that, under general conditions, the matrix of observed indicators is less than the matrix of latent scores and errors. Therefore, it can be demonstrated that even if the model is identified the latent scores are indeterminate. There are infinite sets of latent scores for the same identified model. It can be proved that some of them can be either negatively correlated to each other (Reiersol 1950; Guttmann 1955; Anderson and Rubin 1956; Lawley and Maxwell 1963; Joreskog 1967; Schonemann and Wang 1972; Schonemann and Steiger 1978; Steiger 1979; Schonemann and Haagen 1987). In this case, given $\mathbf{Q}_{\mathbf{y}_{14}}\,\mathbf{y_{17}}$ and $k_2^{\#}$, we can obtain infinite set of scores of **HC**; moreover some of them can be negatively correlated**.**

An alternative proposal is given by the Partial Least Squares Method (from here on referred to as PLS): PLS provides estimates of parameters **g** in (2) defining and estimating an LV "by deliberate approximation as a linear aggregate of its observed indicators" (Wold 1982). In this definition the **HC** appearing in (2) is not a factor of the observed reflective indicators (3) but an unobserved theoretical construct, approximated by a linear combination of observed formative indicators, e.g. following equation (2):

$$\mathbf{H\hat{C}} = \mathbf{F}\,\hat{\mathbf{g}} \qquad (6)$$

where $\mathbf{H\hat{C}}$ is the proxy obtained by reducing the loss of information with respect to the unobservable HC. There are two alternatives for obtaining the solutions of $\mathbf{H\hat{C}}$ in (6) by means of the PLS. The PLS mode A is based on iterative multivariate regressions of the LV's on the observed indicators; therefore, if there is a single LV, it cannot be used, because it causes "circular solutions" without improvements in the iterations. The PLS mode B is based on simple iterative regressions on the observed indicators $\mathbf{F}=(\mathbf{y_{14}};\ \mathbf{\Psi})$. It can be proved that the estimate of $\mathbf{H\hat{C}}$ is equivalent to the first principal component of **F** (Wold 1982). Therefore we have in (6):

$$\mathbf{H\hat{C}} = \mathbf{F}\mathbf{v_1} = \mathbf{y_{14}}\, v_{11} + \mathbf{\Psi}\ \mathbf{v_{12}} \qquad (7)$$

where $\mathbf{F} = (\mathbf{y_{14}}, \mathbf{\Psi})$ and $\mathbf{v_1}' = (v_{11}, \mathbf{v_{12}})'$ is the first eigenvector of $\mathbf{F}'\mathbf{F}$, $\mathbf{H\hat{C}}$ is the first principal component of $\mathbf{F}$ after its standardization to unit variance ($Var(\mathbf{H\hat{C}})=1$), v_{11} contains the element of the first eigenvector connected with $\mathbf{y_{14}}$, $\mathbf{v_{12}}$ is the sub-vector of $\mathbf{v_1}$ connected with $\mathbf{\Psi}$.
First of all, the estimate $\mathbf{H\hat{C}}$ of $\mathbf{HC}$ does not take into account the actual effects of the investment in $\mathbf{HC}$ on the income and wealth of the households.
Secondly, also from the statistical point of view, there are some general critique about the solutions obtained by means of PLS (Garthwaite 1994).
In this case, in particular, every solution that can be obtained in (3) starting from (7) is logically inconsistent. In effect, substituting $\mathbf{HC\hat{}}$ obtained by (7) in (4) we obtain:

$$\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{y_{17}} = \mathbf{H\hat{C}}\ k_2^{\#} + \mathbf{u_{17}} \qquad (8)$$

and from (7) we have:

$$k_2^{\#} = \mathbf{HC\hat{}}'\mathbf{Q}_{\mathbf{y}_{14}}\mathbf{y_{17}} = (\mathbf{y_{14}}\, v_{11} + \mathbf{\Psi}\,\mathbf{v_{12}})'\mathbf{Q}_{\mathbf{y}_{14}}\mathbf{y_{17}}$$
$$= \mathbf{v_{12}}'\,\mathbf{\Psi}'\mathbf{Q}_{\mathbf{y}_{14}}\mathbf{y_{17}} \qquad (9)$$

from which $k_2^{\#}$ cannot consider the whole $\mathbf{HC}$ contribution to earned income $\mathbf{Q}_{\mathbf{y}_{14}}\mathbf{y_{17}}$ because, by definition, the indirect contribution of Wealth on Income $\mathbf{y_{17}}$ by means of $\mathbf{H\hat{C}}$ is null.
However, if we consider equation (3) where the dependent variable is Income $\mathbf{y_{17}}$ we have, substituting $\mathbf{H\hat{C}}$ obtained by (7):

$$\mathbf{y_{17}} = [\mathbf{y_{14}}, \mathbf{y_{14}}, \mathbf{\Psi}] \begin{pmatrix} 1 & 0 \\ 0 & v_{11} \\ 0 & v_{12} \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} + \mathbf{u}_{17} \qquad (10)$$

In (10) we observe the presence of collinearity between regressors, and if we join the parameters, we cannot divide the direct contribution of Invested Wealth on Income and the indirect contribution of Wealth by means of $\mathbf{H\hat{C}}$. In effect,

$$\mathbf{y_{17}} = \mathbf{y_{14}}\,[k_1 + v_{11}\,k_2] + \mathbf{\Psi}\mathbf{v_{12}}\,k_2 + \mathbf{u_{17}} \qquad (11)$$

The Latent Variable Approach: a new proposal
It has been shown that the solutions obtained by means of the Factor Model are not unique and that the solutions obtained by the PLS Method are not logically consistent (Lovaglio 2003). In order to overcome this problem, a solution can be found in the use of all the information embedded in the Path Analysis model (2) (3). In this way, the $\mathbf{HC}$ is not previously obtained in equation (3) but, respecting the economic relationships is simultaneously obtained from reflective and formative indicators. In this perspective, observing the Path Analysis model (2) and (3), $\mathbf{HC}$ can be defined as a multidimensional construct approximated by the linear combination of its formative indicators ($\mathbf{y_{14}}$,$\mathbf{\Psi}$) that better fits the only reflective indicator $\mathbf{Q}_{\mathbf{y}_{14}}\mathbf{y_{17}}$, that we can define as the earned income effect.
Therefore we have from (2) :

$$\mathbf{Q}_{\mathbf{y}_{14}}\mathbf{y_{17}} = \mathbf{F}\mathbf{g}k_2 + \mathbf{u_{17}} = \mathbf{F}\,\mathbf{k_3} + \mathbf{u_{17}} \quad \text{where } \mathbf{k_3} = \mathbf{g}k_2 \qquad (12)$$

In (12) we obtain $\mathbf{k}_3^*$ by means of an ordinary Least Squares Regression of $\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{y_{17}}$ on $\mathbf{F}$. The $\mathbf{k}_3^*$ vector contains the effects of the formative indicators $\mathbf{F}$ on earned income $\mathbf{Q}_{\mathbf{y}_{14}}\mathbf{y_{17}}$:

$$\mathbf{k}_3^* = \mathbf{g}\,k_2 = \mathbf{S_F}^{-1}\mathbf{F}'\mathbf{Q}_{\mathbf{y}_{14}}\mathbf{y_{17}} \quad \text{where } \mathbf{S_F} = \mathbf{F}'\mathbf{F} \qquad (13)$$

Premultiplying the equation (13) by $\mathbf{F}$ and taking into account (2) we obtain:

$$\mathbf{F}\,\mathbf{k}_3^* = \mathbf{F}\,\mathbf{g}\,k_2 = \mathbf{HC}\,k_2 \qquad (14)$$

Remembering that $Var(\mathbf{HC}) = \mathbf{S_{HC}} = 1$ we reach:

$$\mathbf{k}_3^{*\prime}\mathbf{S_F}\,\mathbf{k}_3^* = k_2\,\mathbf{S_{HC}}\,k_2 = k_2^2 \qquad (15)$$

From (15) we obtain k_2^*, the effect of $\mathbf{HC}$ on income net of wealth $\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{y_{17}}$:

$$k_2^* = [(\mathbf{y_{17}}'\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{F}\,\mathbf{S_F}^{-1}\mathbf{F}'\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{y_{17}}]^{1/2} =$$
$$= [\mathbf{y_{17}}'\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{P_F}\ \mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{y_{17}}]^{1/2} \qquad (16)$$

where $\mathbf{P_F} = \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$.

Therefore, from (13) and (16), we obtain $\mathbf{g^*}$, the effect of the formative indicators $\mathbf{F}$ on $\mathbf{HC}$:

$$\mathbf{g^*} = \mathbf{k}_3^* / k_2^* =$$
$$= [\mathbf{y_{17}}'\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{P_F}\,\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{y_{17}}]^{-1/2}\ \mathbf{S_F}^{-1}\mathbf{F}'\mathbf{Q}_{\mathbf{y}_{14}}\ \mathbf{y_{17}} \qquad (17)$$

At this point from (2) and (17) we obtain the estimation of $\mathbf{HC}$ scores ($\mathbf{HC^*}$) :

$$\mathbf{HC^*} = \mathbf{F}\,\mathbf{g^*} \qquad (18)$$

The Latent Variable Approach: mixed indicators
In our case, some of the formative indicators are categorical.
Therefore we partition the vector of formative indicators into quantitative (contained in the column of matrix $\mathbf{F_q}$) and categorical indicators $\mathbf{F_c}$ in order to obtain consistent solutions with the quantitative case. We express the equation (2) in the following way:

$$\mathbf{HC} = \mathbf{F_c}\,\mathbf{g_c} + \mathbf{F_q}\,\mathbf{g_q} + \mathbf{u_{16}}, \quad (\mathbf{F_c} = \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}, \mathbf{x_7}) \qquad (19)$$

where $\mathbf{F} = (\mathbf{F_c}, \mathbf{F_q})$, $\mathbf{g} = (\mathbf{g_c}, \mathbf{g_q})$

We avoid the approach of the Item Factor Model (Christoffersson 1975; Olsson 1979; Muthen and Christoffersson 1981) consisting of a Factor Model with qualitative and categorical indicators. In effect, this model increases the difficulties concerning the original factor model. The non realistic hypothesis of the normality of qualitative indicators, and some restrictive assumptions, determine an underestimation of the true correlation between different indicators, the asymptotical distortions of standard errors and the non chi-square distribution of goodness of fit statistics (Quiroga 1991; Vittadini 1999).

We choose the multidimensional scaling method ALSOS, which alternatevely estimates, in separate steps the parameter vector **g** and quantifies the categorical indicators $\mathbf{F_c}$ by means of a unique algorithm., inside a specified model, the Multiple Regression Analysis (De Leeuw, Young and Takane 1976; De Leeuw and Young 1978; De Leeuw and Van Rijckevorsel 1980; Young 1981; Gifi 1981; Keller and Waansbeek 1983).

We adapt this methodology in order to obtain the **HC*,** with the same methodology proposed in the quantitative case. If at the first step we arbitrarily choose the parameters $\mathbf{g_c^{(0)*}}$, we obtain the first quantification of $\mathbf{F_c^{(0)*}}$ $\mathbf{F_c}$

$$\mathbf{F_c^{(0)*}} = \mathbf{F_c} \quad \mathbf{g_c^{(0)*}} \text{ with } \mathbf{F^{(0)*}} = (\mathbf{F_c^{(0*)}}, \mathbf{F_q}) \qquad (20)$$

At this point, we introduce $\mathbf{F^{(0)*}}$ in (11) using equations (13)-(17), we obtain the first estimates of the parameters $k_2^{(1)*}$, $\mathbf{k_3^{(1)*}}$, $\mathbf{g^{(1)*}} = (\mathbf{g_c^{(1)*}}, \mathbf{g_q^{*}})$ and by means of (18) the first estimates of **HC, HC** $^{(1)*}$. Using $\mathbf{g_c^{(1)}}$ in (19) we obtain a new quantification $\mathbf{F_c^{(1)}}$ of indicators $\mathbf{F_c}$. The iterative process continues until we have no more changes in k_2^{*} , $\mathbf{k_3^{*}}$, $\mathbf{g^{*}}$, **HC***, **F*** Therefore, in this way, the case of mixed indicators is treated similarly to the case of quantitative indicators.

Human Capital in monetary units.

As a quantitative multidimensional construct, the proposed methodology estimates **HC*** by a linear combination of mixed formative indicators ($\mathbf{y_{14}}$,**Ψ**) that best fits the reflective indicators $\mathbf{Q_{y_{14}}}$ $\mathbf{y_{17}}$. Its estimation is consistent with well established economic theory.

Using the 1983 Federal Reserve Survey of 4,103 households as a representative stratified sample of 83,422,111 American households (Avery and Elliehausen 1985) we obtain the **HC*** scores and distribution which represent the estimated **HC** standardized scores and distribution of the American households (Figure 2).

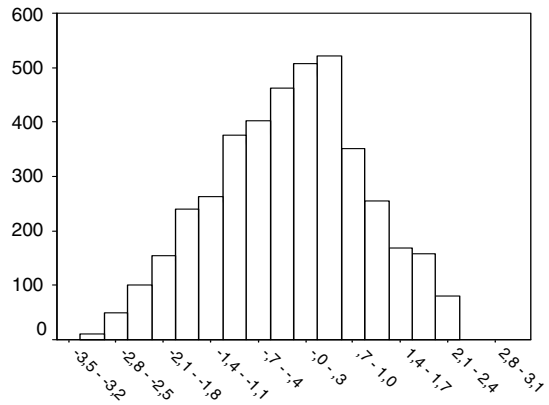

Figure 2: Standardized distribution of HC

At this point, as shown in Dagum and Slottje (2000), we transform the estimated standardized latent variable **HC*** into an accounting monetary value applying the following transformation

$$\mathbf{HC}^{\circ}(i) = \exp [\mathbf{HC^*}(i)] \qquad (21)$$

Then we estimate its mean value :

$$\mu(\mathbf{HC}^{\circ}) = \sum_{i=1}^{n} \mathbf{HC}^{\circ}(i)\, f(i) \,/\, \sum_{i=1}^{n} f(i) \qquad (22)$$

where f(i) is the number of households in the entire population of American households that the i-th sampled household represents, **HC**°(i) is the accounting monetary value of HC*(i) and n is the sample size.

By an actuarial approach the authors estimate the real **HC** upon the idea that an individual's expected mean income at age x+t of a person of age x should be equal to the mean earned income of individuals being at the present x+t years old; therefore the average human capital h(x) of households head of age x is equal to the average expected earned income by age of the households head actualised at a given discount rate and weighted by the survival probability. Hence, the average human capital h(x) of the households head of age x (assumed to stay in the labour market until age 70) is:

$$\mathbf{h(x)} = \Sigma_t \; y_{x+t}\, p_{x,\,x+t}\, (1+ i)^{-t} \quad t = 0,..70\text{-}x \qquad (23)$$

where y_{x+t} is the mean income (real) of the households head of age x + t, $p_{x,x+t}$ is the probability of survival at age x+t of a person of age x and, i is the discount rate (estimated to be 0.08). Therefore the estimation of the average **HC** of the population of American families in monetary units was obtained by Dagum and Slottje (2000) as the weighted mean of **h(x)**:

$$\mu(\mathbf{h}) = \sum_{x=20}^{70} \mathbf{h(x)} f(x) \,/\, \sum_{x=20}^{70} f(x) \qquad (24)$$

The value of average **HC** of the population of American families is estimated to be 238,703$ and it is used to obtain the exponential transformation **HC**° of the standardized latent variable **HC*** in current monetary value. Multiplying **HC**°(i) by the ratio between its mean value μ(**HC**°) and μ(**h**)

$$\mathbf{HC}(i) = \mathbf{HC}^{\circ}(i)\ \mu(\mathbf{h}) / \mu(\mathbf{HC}^{\circ}) \qquad (25)$$

we obtain the vector **HC**(i) of the sample observations in national monetary units, with real mean and variance. The distribution of **HC**(i), which is the estimate of the distribution of **HC** for the entire population of American families in 1983, is plotted in Figure 3.

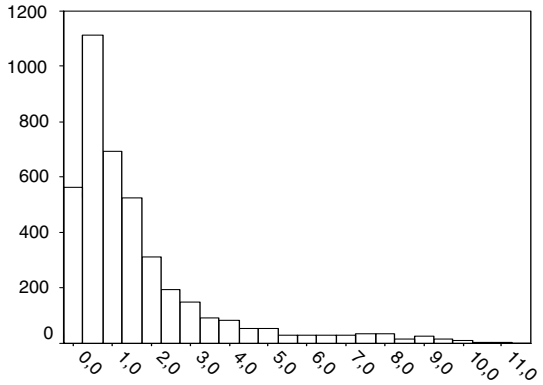

Figure 3: Distribution of US Household (10.000$)

The advantages of the proposed model

The proposed method has several advantages.

1) It uniquely estimates the scores of **HC** from manifest variables (MV), consistently with the supposed causal relations, avoiding treating the formative as reflective and viceversa.

2) The parameters are estimated in a causal model framework because the LV is not exactly defined as a linear combination of its manifest indicators and the error matrix is interpretable as true stochastic errors.

3) The approach is nonparametric.

4) In the case of many dependent reflective indicators this method can be generalized by means of the Redundancy Analysis (Tso, 1981), proposed in PLS Path modeling (Tenenhaus, 1995; Lovaglio, 2001).

References

Anderson, T. W. and Rubin, M. (1956). Statistical inference in factor analysis. *Proceeding of the third Berkeley Symposium on mathematical statistics and probability* 5, pp.11-150.

Avery, R. B. and Elliehausen, G. E. (1985). *1983 Survey of Consumer Finances: Technical Manual and Codebook*, Federal Reserve Board, Washington.

Becker, G. S. (1962) Investment in Human Capital : a Theoretical Analysis, *Journal of Political Economy*, vol. LXX, n.5, Part 2, pp.9-49.

Becker, G. S. (1964). Human Capital, Columbia University Press and NBER, New York.

Bekker, P.A. (1989) Identification in Restricted Factor Models, and the Evaluation of Rank Conditions, *Journal of Econometrics* 41, pp.5-16.

Bekker, P.A. and Pollock D.S.G. (1986) Identification of Linear Stochastic Models with Covariance Restrictions, *Journal of Econometrics* 31, pp.179-208.

Bentler, P.M. (1982) Linear system with multiple levels and types of latent variables. In *System under indirect observation*, Jöreskog K. and Wold H. (eds.), North Holland, Amsterdam, pp 101-130.

Christoffersson, A. (1975). Factor analysis of dichotomized variables, *Psychometrika*, 40, pp. 5-32.

Dagum, C. and Slottje D. J., (2000). A new method to estimate the level of and distribution of household human capital with application, *Structural change and economic dynamics*, 11, pp. 67-94.

Dagum, C. and Vittadini, G. (1996). Human Capital Measurement and Distribution, *Proceedings of the 156th Meeting of the American Statistical Association*, Business and Economic Statistics Section, pp. 194-199.

Dagum, C., (1994). Human Capital, Income and Wealth Distribution Models and Their Applications to the USA, *Proceedings of the 154th Meeting of the American,* pp 253-258.

Dagum, C., Vittadini, G., Costa, M. and Lovaglio, P. (2003). A Multiequational Recursive model of human capital, income and wealth of households with application, *2003 Proceedings of the American Statistical Association*, Business and Economic statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

de Leeuw, J. and Van Rijckevorsel, J. (1980). Homals and Princals, Some Generalizations of Components Analysis, *Data Analysis and Informatics*, pp. 231-241.

de Leeuw, J. and Young, F. (1978). The Principal Component of Mixed Measurement Level Multivariate Data: an Alternating Least Squares Method with Optimal Scaling Features, *Psychometrika*, 43, pp. 279-281.

de Leeuw, J., Young, F. and Takane, Y. (1976). Regression with Qualitative and Quantitative Variables: an Alternating Least Squares with Optimal Scaling Features, *Psychometrika*, 41, pp. 505-529.

Eisner, R. (1985), The Total Incomes System of Accounts, *Survey of Current Business*, January.

Everitt B.S. (1984) An introduction to latent variable models, Chapman and Hall, N.Y.

Fisher, F. (1976) *The Identification Problem In Economics*, Huntingdon, NY: Krieger Publ.

Fuller W.A. (1987) *Measurement Error Models*. Wiley, NewYork.

Garthwaite, P.H. (1994). An interpretation of Partial

Least Squares, *Journal of American Statistical Association*, 89, pp. 122-127.

Geraci V. (1976) Identification of Simultaneous Equation Models with Measurement Error, *Journal of Econometrics* 4, 263-283.

Gifi, A. (1981). *Non linear Multivariate Analysis*, Department of data Theory, University of Leiden, The Netherlands.

Guttmann, L (1955). The determinacy of factor scores matrices with implications for five other basic problems of common factor theory, *British Journal of Statistical Psychology*, 8, pp. 65-81.

Joreskog, K. (1967). Some contributions to maximum likelihood factor analysis, *Psychometrika*, 32, pp. 443-482.

Jöreskog, K. G. (1973) A general method for estimating a linear structural equation system, in Goldeberger A. S. and Duncan O. D. (Eds.), *Structural equation models in the social sciences,* Seminar Press, New York, 85-112.

Jöreskog, K. G. (1981). Analysis of covariance structures, *Scandinavian Journal of Sstatistics*, 8, 65-91.

Jöreskog, K. G. (1982). The Lisrel approach to causal model building in the social sciences. In K. G. Jöreskog, & H. Wold (Eds.). Sustems under indirect observation *Causality structure prediction Vol 4* (pp81-99). New York: North Holland

Jorgenson, D. M. and Fraumeni, B. M. (1989). The accumulation of human and non human capital, 1948-84. In Lipsey, R. E., Stone Tice, H. (Eds), *The Measurement of Saving, Investment, and Wealth*, *NBER Studies in Income and Wealth*, University of Chicago Press, Chicago, 53, pp. 227-282.

Keller, W.J. and Waansbeek, T. (1983). Multivariate methods for quantitative and qualitative data, *Journal of Econometrics*, 22, pp. 91-111.

Kendrick, J. W. (1976). The formation and Stocks of Total Capital, Columbia University Press, New York.

Lawley, D.N and Maxwell, A. E. (1963). *Factor analysis as a statistical method*. London: Buttherworths.

Lovaglio, P.G. (2001). The estimate of latent outcomes in *Proceedings on Processes and Statistical Methods of evaluation*, Scientific Meeting of Italian Statistic Society, Rome, Tirrenia, pp. 393-396.

Lovaglio, P.G (2003). The estimate of customer satisfaction in a reduced rank regression framework, *Total Quality Management*, in press.

Mincer, J., (1958).Investiment in Human Capital and Personal Income Distribution, *Journal of Political Economy,* vol. LXVI, pp. 281-302.

Mincer, J., (1970).The distribution of Labor Income: A Survey, *Journal of Economic Literature*, vol. VII, n.1, pp.281-302.

Muthén B. and Christoffersson A., (1981). Simultaneous factor analysis of dichotomous variables in several groups, *Psychometrika*, 46, pp. 407-419.

Olsson, U. (1979) Maximum Likelihood Estimation of Polychoric Correlation Coefficient, *Psychometrika*, 44, pp. 443-460.

Quiroga A., (1991). Studies of the polychoric correlation and another measures for ordinal variables. *PhD Dissertation*, Uppsala University, Department of Statistics.

Reiersol, O. (1950). On the identifiability of parameters in Thurstone's multiple factor analysis. *Psychometrika*, 15, pp.121-149.

Reiersol, O., (1950). On the identificability of parameters in Thurstone's multiple factor analysis, *Psychometrika* 15 121-149.

Rothenberg, T.J, (1971) Identification in Parametric Models, *Econometrica*, Vol. 39 pp. 577-591.

Schonemann, P., and Steiger, J.H (1978). On the validity of indeterminate factor scores . *Bulletin of the Psychonomic Society*, 12, pp. 287- 290.

Schonemann, P., and Haagen, K. (1987). On the use of factor scores for prediction. *Biometrical Journal* 29, pp.835-847

Schonemann, P., and Wang, M. M. (1972). Some new results on factor indeterminacy, *Psychometrika* 37, pp. 61-91

Schultz, T. W. (1959), Investiment in Man: An Econominist's View, *Social Science Rewiew*, vol.XXXIII, n. 2, pp. 109-117.

Schultz, T. W. (1961). Investiment in Human Capital, American Economic Rewiew, vol.LI, n.1, pp.1-17.

Shapiro, A. (1985) Identifiability of Factor Analysis: Some Results and Open Problems, *Linerar Algebra and its Application*, 70, pp.1-7.

Steiger, J H (1979). The relationship between external variables and common factors, *Psychometrika,* 44 pp. 93-97

Tenenhaus, M. (1995). *La Régression PLS: Théorie et Pratique*. Editions Technip, Paris..

Tso, M.K.S., (1981). Reduced Rank Regression and Canonical Analysis, *Journal of the Royal Statistical Society*, Series B, 43, pp.183-189.

Vittadini, G. (1999). Analysis of qualitative variables in Structural Models with unique solutions. In: M. Vichi, O. Opitz, (eds.), *Classifications and data analysis-Theory and Applications*, Springer and Verlag, pp. 203-210.

Vittadini, G. and Lovaglio, P.G. (2001). The estimate of latent variables in a structural model: an alternative approach to PLS. In *PLS and Related Methods*. *Proceedings of the PLS International Symposium*. CISIA CERESTA, Montreuil, France, 423-434.

Wegge L. (1965) Identifiability Criteria for a System of Equations as a Whole, *The Australian Journal of Statistics,* 7, pp.67-77.

Wegge L. (1991), Identification with Latent Variables, *Statistica Neerlandica,* 45, 2, pp121-144.

Wegge L. and Feldman M. (1983), Identifiability Criteria for Muth-Rational Expectations Models, *Journal of Econometrics,* 21, pp. 245-254.

Wold, H. (1982). Soft modelling: the basic design and some extension. In: K. Joreskog and Wold H. (eds.), *System under indirect observation*, North Holland, Amsterdam, pp. 1-53.

Young, F. (1981). Quantitative Analysis of Qualitative data, *Psychometrika*, 46, pp. 357-388.