Identifying the Poor:
A Multiple Indicator Approach

Ramses M Abul Naga
London School of Economics and Political Science

Discussion Paper
No.DARP/9
June 1994

The Toyota Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020-7955 6678

I am grateful for advice from Tony Atkinson, Frank Cowell, Steve Howes, Georg Inderst, Magda Mercader and Dirk van de Gaer.

Abstract

The standard approach to the study of poverty assumes the existence of an ideal variable that captures the extent of deprivation. In this paper we postulate that poverty is involved with many dimensions. We use a latent variable framework to predict the extent of an individual's hardship as a function of $\psi_i = ax_{1i} + bx_{2i} + \ldots$, where the x's are indicators of I's income status, yi, and latter variable is not observed.

Keywords: Poverty, latent variable, indicators.

JEL Nos.: I32, C39.

1 Introduction

Two important aspects of any empirical study of poverty are what are known as the identification and aggregation issues. The identification rule dictates how we decide who is poor and often, but not always, it deals with the question of how poor the person is. The aggregation step enables us then to take the individual poverty data, and to summarize them into an economy-wide measure of poverty. For instance when we use the head-count to study the incidence of poverty, we label a person as poor if he falls under the poverty line. The head-count measure is the proportion of the population that is below the specified poverty line.

In this paper we will be concerned with the identification problem in face of imperfect information. Assume that underlying our study of poverty lies a social welfare function $W(y_1,...y_n; y^*)$ where y_i denotes the income status of family i, and y^* is the poverty line. The identification rule then consists of separating families into two groups. The group for which $y<y^*$ is that of the poor. The families for which $y\geq y^*$ are members of the second group; the non-poor . The problem we are dealing with in this work is that of the identification of the poor when y is not observed, or is subject to measurement error. The income status of a family may be measured with error because it is systematically under-reported. If families may benefit from various welfare programmes when their incomes are low, they may have an incentive to understate their resources. Glewwe(1990) adopts this point of view in his work on the "efficient allocation of transfers to the poor". There are of course other reasons why y may be unobservable in practice. If the pertinent variable to welfare analysis is the long run economic status of the family (e.g. its permanent income), then one must recognize that static indicators such as current income and consumption, may be noisy correlates of long run income status. In our work we will adopt this long run income, interpretation of y.

When y is unobserved, one may attempt to isolate its least noisy indicator, as Anand and Harris (1991) and Chaudhuri and Ravallion (1992) have done. However, one cannot ignore the fact that the most commonly used indicators of well-being, namely income and consumption, often offer different information about the composition of the poor: some families may cross the poverty line in the income space but not in the consumption space, and vice versa. It is not clear though why we should not simultaneously exploit

information about various indicators of welfare. If, let us say, consumption is a good indicator of permanent income, is it reasonable to assume that it exhausts all the necessary information about the latter variable?

The aim of this paper can be stated as follows: starting from a set X of p observed indicators on the family's socio-economic characteristics (for example: income, consumption, and employment status of head), we want to construct a summary statistic of the family's long run income status y . The task of constructing such an index can be stated in the language of multivariate statistics, as one of reducing the dimensionality of the data from p to one dimension. As this is not a new problem in statistics, we have a vast literature and many results to draw from.

The index we select must achieve the reduction of the dimension of the data, without losing any available information contained in X, about y. That is, the index must be a *sufficient statistic* for y. One such statistic we propose to use is the regression function of y conditional upon X. Below, we will refer to it as the *multiple indicator index.*

To arrive at the multiple indicator index, we will proceed in the following steps. In section 2 we discuss why the two rival indicators of welfare, namely income and consumption, are not likely to rank families in terms of their well-being in an identical fashion. We also take a first look at our income and consumption data , and note the existence of conflicting conclusions offered by the two indicators. In section 3, we postulate the statistical relation between the indicators X and the unobserved income status of the family, y. We assume that the correlation between the various indicators arises out of their common dependence on y. This formulation leads us to adopt the model of *factor analysis* in describing the relation between the indicators and the unobservable. Within the context of a simplified example of the life-cycle model, we discuss the inferential problems involved with the factor analysis model. We then show how arrive at the multiple indicator index, which enables us to summarize the information about y.

In section 4 we illustrate the multiple indicator approach using a sample of 910 families, extracted from wave XX of the Panel Study of Income Dynamics. We then construct two classification matrices, where observations are ranked using the multiple indicator index, and each of the two commonly

used indicators; consumption and income. These matrices offer greater scope for agreement concerning the ranking of families, than in the case where observations are ranked on the basis of income and consumption. We therefore summarize the results obtained from the suggested method, as being a "middle of the road" solution as an alternative to working with either of the two indicator separately. In section 5 we comment further on the use of the multiple indicator method, especially with respect to the limitations of its applicability. Section 6 concludes the paper.

2 The Problem

As a first step to any study on the extent of poverty we have to specify the choice of our indicator variable of welfare. Let x denote such an indicator, and x^* the specified poverty line. Should we define the standard as an income measure x_1 or should it be expenditure x_2? Conflicting conclusions can arise when one indicator is chosen at the expense of another. Consider first the case when $x_1 < x_1^*$ and $x_2 > x_2^*$. In Crime and Punishment, Dostoyevsky writes the following:

"... possibly what weighed with her most was 'the poor man's pride' which makes poor people who are faced with the necessity of observing certain of our traditional customs strain every nerve and spend their last savings so that they should be 'as good as everybody else' and that no one 'could have a wrong word to say against them'".

Thus, a family may be able to reach a decent level of expenditure by temporarily borrowing or running down its assets. On the other hand, a situation where $x_1 > x_1^*$ and $x_2 < x_2^*$ can occur when there exist market imperfections due to information problems, discrimination, and other obstacles to trade. A well-off black family may afford a rent in a white suburb, but may be obliged to live on the other side of the railroad tracks due to the reasons mentioned above.

The need for examining multiple indicators of welfare has long been recognized in development economics. For instance, in their study of economic mobility and agricultural labour in Rural India, Dreze et al.(1992) write the following:

"One may however question whether current per-capita income in any particular year is a sensible criterion of 'poverty' in economies where current incomes are subject to large short-run

variations and significant mechanisms exist for smoothing out these fluctuations. On the basis of alternative criteria of poverty such as per-capita expenditure or living standard, it is likely that less mobility would be observed. Households which may appear to be "crossing the poverty line " in particular years in the income space may, in fact, be chronically poor in terms of expenditure or living standards."

Also, Glewwe and Van der Gaag (1990) test on L.D.C. data the consequences of using various definitions of deprivation in identifying the poor. They conclude that the choice of a resource variable matters significantly, in the sense that different definitions identify different groups of the population as poor.

It is worthwhile asking the question as to when one can safely use one indicator instead of another; i.e., under what circumstances is it that the problem of selecting a specific indicator amongst several, ceases to be a "problem". Let us once more restrict ourselves to the case where the choice is between two indicators: income (x_1), and consumption (x_2). We also abstract from the borrowing and other quantity constraints mentioned earlier. Then,

(i) if the correlation between x_1 and x_2 is equal to unity, one can use the information on either indicator to identify the poor. For example, if individuals all face the same prices, the information contained in nominal and real incomes will be identical to the researcher.

(ii) one could weaken the condition (i) , in requiring that the ordering of individuals by x_1 and x_2 be preserved. The choice between x_1 and x_2 would be unimportant provided there were a one-to-one relation between the two variables. One could then also convert the poverty line in terms of x_1 into a poverty line in terms of x_2 .

The first economic example that comes to our minds , is that of the Keynesian consumption function. With a simple linear relation of the type $x_2 = a+bx_1$, the income poverty line x^*_1 is mapped into a consumption poverty line $x^*_2=a+bx^*_1$, leaving the identification of the poor invariant to the choice of indicator.

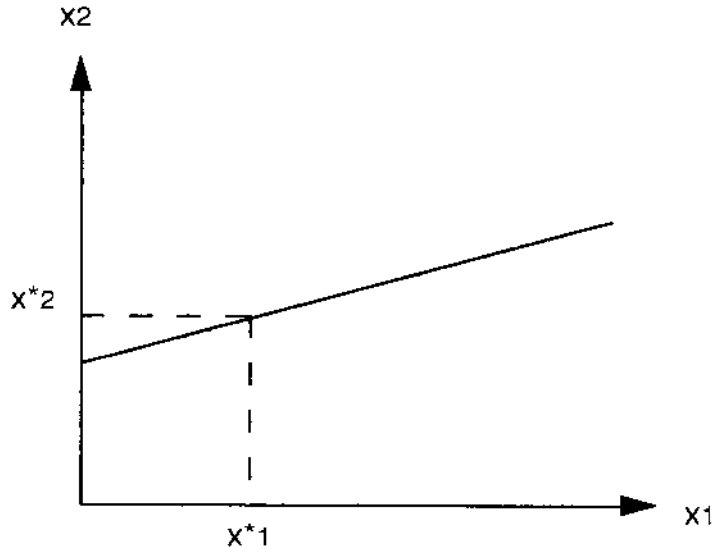

FIGURE 1: Income and consumption poverty lines.

The linear consumption function is not likely to be of much relevance at the microeconomic level, unless we have a good reason to assume that savings do not respond to the interest rate. As mentioned above, we need not restrict ourselves to a linear relation between income and consumption, any one-to-one relation will serve our purposes. Deaton(1992) ch. 1, offers several reasons why one cannot expect consumption to closely follow income. One explanation relates to the fact that the marginal utility of consumption may vary along the stages of the life-cycle, especially with respect to the demographic structure of the household. More importantly, when we incorporate uncertainty in the course of the life-cycle, we note that the consumption patterns of households may substantially vary because of their different abilities to bear risk. As Deaton (p.19) very well summarizes the fact:

"In an uncertain world, the substitution of future consumption for current consumption inevitably increases exposure to risk, and those who are willing to contemplate the former must be willing to face the latter."

Life-cycle income would then be a key variable in the ability to bear risk. Looking at the relationship between current income and consumption only, would leave us with an incomplete, and possibly, inaccurate picture. A priori grounds therefore, we cannot assume a one-to-one relation between consumption and income. We would thus expect a substantial reranking of families according to which of the two indicators we use.

At this stage, it is worth confronting the evidence by taking a first look at the data we will be making use of in this work. We have data on income to needs ratio and food budget share for 910 American families. These observations were extracted from the Panel Study of Income Dynamics, and pertain to 1986. Ideally we would want to use an expenditure to needs ratio for the consumption indicator. However the PSID does not possess as refined data on consumption as the ones available on income. Noting that food consumption is typically an inferior good (at least in developed countries), a negative correlation between the income to needs ratio and food budget share is to be expected. In our data, It is of the order of -0.3 . In the matrix below, we classify our observations line-wise in order of increasing income , and column-wise in decreasing order of food budget share. We adopt a quartile classification, so that income and consumption classes are of equal size, and contain exactly one quarter of the families.

$$T(x_1, x_2) = \begin{bmatrix} 138 & 51 & 23 & 15 \\ 65 & 88 & 51 & 24 \\ 20 & 66 & 86 & 55 \\ 5 & 22 & 68 & 113 \end{bmatrix}$$

By examining the off-diagonal elements of the matrix, one notes the extent of divergence in the classification of observations by the two indicators. A χ^2 test of independence between x_1 and x_2 is quite expectedly rejected at 99%. While the indicators do not rank the families identically, there is a correlation between classifications obtained by the two variables (this result is not surprising since the correlation between x_1 and x_2 is -0.3). The extent of disagreement between the two indicators in the ranking of families can be quantified by counting the ratio of off-diagonal elements to the total number of observations. There are 465 out of 910 families ranked differently by x_1 and x_2, thus over 50% of the total. By setting the poverty line at a particular level for each indicator, we may of course observe that agreement between the two indicators may be higher or lower for this subset of individuals, than for the entire population. On the other hand, there is no general agreement on how, and at what level to set the poverty line. It would therefore appear more reasonable to discuss the adequacy of specific indicators in the applied analysis of living standards independently of where we may wish to set the poverty line. We will thus examine the performance of the indicators on the entire range of the population, rather than limit ourselves to a specific subset .

Nonetheless, for illustrative purposes in the present discussion, assume that the poverty line is set on a relative basis, so that the bottom 25%, ranked according to a specific indicator, are defined as the poor. Then, from our classification matrix T we see that 138 families are ranked poor by both indicators; whereas 227 (i.e. 902/4) by construction, are identified as poor using each of the income and consumption indicators. In 40% of cases, the most popular indicators, income and consumption, identify different families as being in poverty. This result is largely in conformity with many economic theories of consumption. Since income is correlated with consumption, one indeed expects that the classification using the two indicators does not exhibit a pattern of independence. On the other hand, since the relationship between current income and consumption is not one-to-one, the ranking of individuals according to the two indicators is not likely to be uniform. We thus conclude this section by stating that indeed the choice between indicators in identifying the poor is a "problem".

3 A Multiple Indicator Approach

The evidence that alternative indicators offer different information concerning the identification of the poor, can be interpreted in three different ways. The first argument would be to say that there exists a proper indicator of welfare, say income, and to postulate that consumption is related to it through an economic law of the type:

$$x_2 = a+bx_1 + \varepsilon$$

where ε is a random term. The researcher is then entitled to prefer x_1 to x_2. This is probably the status-quo amongst researchers today, divided between those who favour the choice of consumption, and those who opt for income. Chaudhuri and Ravallion(1992) write:

"Current consumption expenditure and current income have been the most popular welfare indicators in applied welfare analysis. Of the two, current consumption expenditure has probably been more widely used for research and policy purposes, at least in developing countries."

An alternative approach is to argue that both income and consumption are correlates of an unobserved variable, which is pertinent to welfare analysis. For example, if our main concern is with chronic, rather than transient poverty,

(the concern of Chaudhuri and Ravallion(1992)), one would expect current income and consumption to be noisy measures of long term economic status. This time, the preference for x_1 over x_2 could be motivated on the basis that x_1 is a noisy indicator of a variable y, defined as long term economic status, and as before, x_2 is a noisy indicator of x_1 . The problem arises because y typically is not observable. For example, if y is permanent income, there is no direct way of measuring it . Thus, permanent income is to be treated as a latent variable. Let us write a simple system of equations relating x_1 ,x_2, and y.

$$x_2 = a + bx_1 + w$$

$$x_1 = y + e$$

We see that x_2 will be a noisy indicator of y, but this time, it will be noisier than x_1. Substituting x_1 into the equation for x_2, we get :

$$x_2 = a + by + v$$

where the variance of v equals b^2var(e)+var(w). The error term for x_2 is correlated with that of x_1, and the variance of x_2 conditional on y, is greater than the corresponding one for x_1. In this sense, x_1 is a better indicator of y.

Anand and Harris (1991) essentially adopt the above framework in their analysis of living standards in Sri- Lanka. Under certain conditions they are able to show that if x_i is more variable than x_j, average savings in a given population decile will be lower when families are ranked according to x_i than when they are ranked according to x_j. This result then allows them to perform pair-wise comparisons on the relative variability of some commonly used indicators of living standards.

The third explanation regarding the fact that alternative indicators offer different information with respect to the identification of the poor, is to be found in the assumption that the indicators are jointly correlated through their common dependence on y. This is the approach we pursue in this work. Let $g(x_i|y)$ denote the probability density function of x_i when y is held fixed. We let x_1 and x_2 be related to y through the following system:

$$x_1 = y + u_1$$

$$x_2 = \beta_2 y + u_2 \qquad (3.1)$$

Define y as permanent income, x_1 as current income, and u_1 transitory income. According to Friedman (1957) p.21,

"The permanent component is to be interpreted as reflecting the effect of those factors that the unit regards as determining its capital value or wealth. ... The transitory component is to be interpreted as reflecting all 'other' factors."

We also let x_2 denote current consumption expenditure and u_2 be a disturbance term in the consumption equation. Equations (3.1) defines a pure life-cycle model, where current income and consumption are endogenous variables and are functions of permanent income. The decomposition of observed income, x_1, into permanent and transitory components, dates back to the work of Friedman and Kuznets (1945) on income inequality in the United States. The equation for x_2 was subsequently added by Friedman (1957) in his work on the consumption function. As with current income, observed consumption was decomposed into a permanent component $\beta_2 y$ and transitory element u_2.

The assumption that the correlation between x_1 and x_2 is solely induced by y can be formally stated as follows:

$$g(x_1, x_2 \mid y) = g(x_1 \mid y) \cdot g(x_2 \mid y)$$

That is, when y is held fixed, x_1 and x_2 must be independent. The above condition is known as the axiom of conditional independence. From the stated axiom , it follows that

$$\mathrm{cov}(u_1, u_2) = 0 \qquad (3.2)$$

Note that now the disturbances are uncorrelated, whereas when we assume x_2 is a function of x_1, the error term in the consumption equation is a function of the error term of of the income equation. In contrast to this, in the present set-up, a priori all indicators have a symmetric status with respect to y.

Our question is the following: can one use the information available from the two indicators, in order to predict the permanent income of the family? The answer hinges upon the identifiability of the model (3.1). The first step thus consists of examining the identification of the model relating the observed variables $X=[x_1,x_2]$, to the unobservable y. If we can estimate all the structural parameters of the model, than, in a second step we can postulate a parametric form for the distribution of the indicators, say g(X|y), from which we may attempt to recover h(y|X). Having observed information on the current income and consumption of a family, we can use the distribution h(y|X) in order to construct an index of the family's permanent income.

Let us first examine the identification of model (3.1). The question of identification can alternatively be stated as follows: from the sample moments available to us, can we identify β_2, and the variance of the two error terms, say ω_{11}, ω_{22} ?

Noting that we have three sample moments var(x_1), var(x_2), and cov(x_1,x_2), we can write the following identities relating sample and population moments:

$$\begin{aligned} \text{var}(x_1) &= \text{var}(y) + \omega_{11} \\ \text{var}(x_2) &= \beta_2^2 \text{var}(y) + \omega_{22} \\ \text{cov}(x_1, x_2) &= \beta_2 \text{var}(y) \end{aligned} \qquad (3.3)$$

To simplify the problem further, assume x_1 and x_2 have been standardized so that var(x_1)=var(x_2)=1. We note from (3.3) that in order to identify the structural parameters of the model, we need to know the variance of y, which we will denote as γ_{yy}. Because γ_{yy} is unknown, the model (3.1) is not identified.

Let us set cov(x_1 x_2) at 0.3, the sample correlation between our income and consumption indicators. In the table below, we consider various values of γ_{yy}, and the corresponding estimates of the structural parameters. Note that in order for us to obtain any meaningful results, the variance of y has to be restricted to the interval $0 < \gamma_{yy} \leq 1$.

γ_{yy}	β_2	ω_{11}	ω_{22}
0.25	1.2	0.75	0.64
0.5	0.6	0.5	0.82
0.75	0.4	0.25	0.88
1	0.3	0	0.91

TABLE 1: Identification of model 3.1 through moment restrictions

The introduction of prior information about γ_{yy} may leave the practitioner uneasy, since different beliefs about the variance of the unobservable y, tend to produce different parameter estimates for the structural coefficients relating the indicators to y .

One thus has to rely on alternative routes to resolving the identification problem (1) . The simplest method is to suppose that panel data are available, say on income x_1. Then, assuming the distribution of income is stationary, we can identify the model with two successive observations on x_1, that we denote below as x_{11} and x_{12}.

$$x_{11} = y + u_{11}$$
$$x_{12} = y + u_{12} \qquad (3.4)$$

From the stationarity assumption, $var(u_{11}) = var(u_{12}) = \omega_{11}$, so that γ_{yy} can also be estimated:

$$var(x_{11}) = \gamma_{yy} + \omega_{11}$$
$$cov(x_{11}, x_{12}) = \gamma_{yy}$$

This is the approach followed by Van Praag et al.(1983) in the measurement of inequality, when income is subject to measurement error.

An alternative way of resolving the identification issue, is to assume the availability of another indicator of y. If y is long term income status, then we can assume that say, asset holdings of the family, are also informative about y. Let x_3 denote this third indicator. The model (3.1), augmented by one extra equation is written below.

$$
\begin{aligned}
x_1 &= \quad y + u_1 \\
x_2 &= \beta_2 y + u_2 \\
x_3 &= \beta_3 y + u_3
\end{aligned}
\qquad (3.5)
$$

Maintaining the independence of the error terms, we can write down the moment restrictions relating the structural parameters and sample moments:

$$
\begin{aligned}
\text{var}(x_1) &= \gamma_{yy} + \omega_{11} \\
\text{var}(x_2) &= \gamma_{yy}\,\beta_2^2 + \omega_{22} \\
\text{var}(x_3) &= \gamma_{yy}\,\beta_3^2 + \omega_{33} \\
\text{cov}(x_1, x_2) &= \gamma_{yy}\,\beta_2 \\
\text{cov}(x_1, x_3) &= \gamma_{yy}\,\beta_3 \\
\text{cov}(x_2, x_3) &= \gamma_{yy}\,\beta_2\,\beta_3
\end{aligned}
$$

Let σ_{ij} denote $\text{cov}(x_i, x_j)$, solving the six equations above we obtain the following solutions:

$$\gamma_{yy} = \sigma_{12}.\sigma_{13} / \sigma_{23} \qquad \omega_{11} = \sigma_{11} - \sigma_{12}.\sigma_{13} / \sigma_{23}$$

$$\beta_2 = \sigma_{23} / \sigma_{13} \qquad \omega_{22} = \sigma_{22} - \sigma_{12}.\sigma_{23} / \sigma_{13}$$

$$\beta_3 = \sigma_{23}. / \sigma_{12} \qquad \omega_{33} = \sigma_{33} - \sigma_{23}.\sigma_{13} / \sigma_{12}$$

It is more convenient to set $\gamma_{yy}=1$ and to introduce a slope parameter β_1 for the indicator relating the indicator x_1 to y. Under such normalization β_i becomes the correlation coefficient between x_i and y (and writing the moment equation for x_1, we see that β_1 is equal to γ_{yy}).

The general model relating the indicators X to y is written as follows:

$$
\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{bmatrix}
=
\begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}
y \quad +
\begin{bmatrix} u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_p \end{bmatrix}
$$

or, in a more compact notation:

$$X = \beta y + U \qquad (3.7)$$

where $\beta=[\ \beta_1, \ldots, \beta_p\ ]'$ and $U=[\ u_1, \ldots, u_p\ ]'$. Model (3.7) is known as the *factor analysis* model. We have now informally discussed the identification of the model relating the indicators to the observable long term income status of the family unit [2]. Other than ranking indicators in terms of their correlation with y, it is not clear how the above method might provide helpful guide-lines in identifying the poor. For example assume we estimate the model with three indicators, and say we find that the coefficient on income β_1, is estimated at 0.8, while the coefficient on consumption β_2, is equal to 0.5, and that on assets, β_3, equals 0.2. We may then decide that out of the three indicators, current income x_1 should be selected because of its greatest association with permanent income. The above statement is equivalent to saying that x_1 is the least noisy indicator of y, since the variance of u_i decreases as β_i rises.

As an alternative to selecting the least noisy indicator of y, we may wish to pool the information available from all the indicators in order to predict, in a sense, the permanent income of the families. One can then hope to rank the families in terms of their predicted permanent incomes, and separate them out between poor and non-poor, according to a specified level of the poverty line.

The task of predicting y , having observed X , requires the derivation of the distribution of y conditional on X. At this stage it is necessary to make assumptions regarding the joint distribution of X and y, denoted as f(X,y). Having specified the parametric form of f(X,y), one can use Bayes' rule to arrive at h(y|X):

$$h(y|X) = f(X,y) \ / \int f(X,y)dy$$

In terms of the general model (3.7) relating the observed variables X to the unobservable y, we assume that $U \sim N(0,\Omega)$, and $y \sim N(0,1)$. It follows that

$$X \mid y \sim N(\beta y,\ \Omega) \qquad (3.8)$$

and

$$X \sim N(0\ ;\ \beta\beta' + \Omega) \qquad (3.9)$$

Then we can use some properties of the normal distribution (e.g. see Greene(1991) p.78) to establish that the conditional distribution h(y I X) is also normal:

$$y| X \sim N [\beta' \Sigma^{-1} X ; 1- \beta'\Sigma^{-1} \beta] \qquad (3.10)$$

where $\Sigma = \beta\beta'+\Omega$ is the variance of the vectors of indicators.

We use the mean of h(y|X) as our index of long run income, i.e.

$$E(y|X) = \beta' \Sigma^{-1} X \qquad (3.11)$$

The suggested index is the regression function of y conditional upon X, and is a linear function of the indicators. Before we further discuss the multiple indicator approach, we analyse our data using the suggested method. Let ψ denote E(y|X). Below we refer to ψ as the *multiple indicator index*.

4 Application to U.S. data

We now illustrate the multiple indicator method on the basis of a sample of 910 families, extracted from wave XX of the U.S. based, Panel Study of Income Dynamics. All the families are male headed and the data pertain to 1986. The selection method as well as the main characteristics of the sample are described in Abul Naga (1994) pp.4-7. The point to note here is that the families retained all have a head who is participating in the labour force. Thus, individuals in full time education, elderly families, and other households who do not participate in the labour force are excluded from the analysis. The consequence of this selection rule is that families for which income is low and consumption is (relatively) high, may be under-represented. Nonetheless, since the application here is meant to be illustrative, rather than descriptive, we decided to go ahead with the application of the multiple indicator method on the basis of these 910 families.

In a first stage we report results for the estimation of the model (3.7) relating three indicators, X, to an unobserved welfare standard y. Then we will construct the multiple indicator index E(y|X) of (3.11), and examine the insights it offers into identifying the poor. the X variables are the following:

x_1 : family income to needs ratio.

x_2 : food expenditure / total taxable income of head and wife.

x_3 : total annual employment hours of head / total annual employment+ unemployment hours of family head.

The three indicators can be taken to be correlates of permanent income. We would expect β_1 and β_3 to be positive. Even when leisure is a normal good, we expect β_3 to be positive, since the state of being unemployed is different from that of being out of the labour force altogether. We expect β_2 to be negative, since food consumption is an inferior good. We do not have other sources of information on consumption of non-durables in the PSID other than the annual household expenditure of the family on food.

The sample correlation matrix is the following:

$$S = \begin{bmatrix} 1 & & \\ -0.3005 & 1 & \\ 0.2099 & -0.2271 & 1 \end{bmatrix}$$

Parameter estimates of the model $X = \beta y + U$ are reported in table 2 below:

parameter	β_1	β_2	β_3	ω_{11}	ω_{22}	ω_{33}
estimate	0.527	-0.570	0.398	0.722	0.675	0.841
t-value	9.238	-9.475	8.250	11.857	10.045	17.197

Table 2: Parameters estimates and t-values

The coefficient of determination, which is defined as 1 - det $(\hat{\Omega})$ / det (S) takes the value 0.513.

Our coefficients are of the predicted signs. We note also that the income to needs ratio, and x_2, the consumption to taxable income ratio are of similar magnitude. Though x_2 would appear to be a better indicator, a test that $\beta_1 + \beta_2 = 0$, is not rejected at 95%. The difference in magnitude between the two parameter estimates is therefore not significant. The coefficient for the employment ratio β_3 , is estimated to be about 0.4. While the coefficient of

determination is not very impressive, we should note that the coefficients are all highly significant.

Having estimated the structural parameters of the model, we can now derive our multiple indicator index ψ , on the basis of (3.11):

$$\psi = E(y|X) = 0.36x_1 - 0.41x_2 + 0.23x_3$$

We note that the budget share of food in this first application is assigned the highest weight. It is followed in importance by the income variable. The variable that carries least weight is the employment indicator of the family head.

In order to assess the impact of using ψ in the identification of the poor, we construct a 4 by 4 classification matrix where families are ranked line-wise by current income x_1, and column-wise using the multiple indicator index ψ:

$$T(x_1,\psi) = \begin{bmatrix} 175 & 52 & 0 & 0 \\ 43 & 144 & 41 & 0 \\ 7 & 31 & 173 & 16 \\ 2 & 1 & 13 & 212 \end{bmatrix}$$

The main diagonal now contains 704 families (that is 77.4% of the families are ranked identically by ψ and x_1). Using income and consumption, agreement was possible only in 445 cases (see the matrix $T(x_1,x_2)$ in section 2). In this sense, wherever we choose to set the poverty line, we are likely to find a closer ranking of families using income and ψ.

Let us now contrast the classification of families using consumption x_2, and the multiple indicator index ψ. We rank families line-wise in order of decreasing food budget share, and column-wise in increasing order of ψ.

$$T(x_2,\psi) = \begin{bmatrix} 169 & 50 & 6 & 3 \\ 39 & 113 & 59 & 16 \\ 13 & 48 & 100 & 67 \\ 6 & 17 & 62 & 142 \end{bmatrix}$$

The main diagonal now contains 524 families, as opposed to the 445 families in the classification by x_1 and x_2. Define d_i as the sum of the diagonal

elements in classification matrix T_i. Also define p as the ratio d/n, where n is the number of observations available in the sample. We can then test the assumption that in matrix T_i agreement is higher than in matrix T_j, using the statistic Z_p :

$$Z_p = \sqrt{n}\,\frac{[\,p_i - p_j]}{\sqrt{p_i(1-p_i) + p_j(1-p_j)}}$$

where Z_p is asymptotically distributed as a N(0,1) variable. In the table below we test the assumption that the two classification matrices constructed by making use of the multiple indicator index offer a wider scope for agreement than in the case of the matrix constructed using the income and consumption indicator

matrix	d	Z_p
$T(x_1,x_2)$	445	-
$T(x_1,\psi)$	704	13.2
$T(x_2,\psi)$	524	3.73

Table 3: Tests of agreement in classification matrices

We see that these differences are all highly significant. Note also in this application that ψ and income offer more agreement than ψ and consumption.

Consider now the following exercise: for various values of the poverty rate, we wish to explore the extent of agreement between income, consumption, and the multiple indicator index ψ. if we decide that say, the lowest ranked 10% are the poor, than we can agree on a maximum of 910/10=91 cases. In table 4 below π denotes the poverty rate, and H is the head count of the poor. Table 4 shows that using ψ with either indicators increases the agreement on the number of poor families. At the lower values of the poverty rate, consumption appears to agree more with ψ, while at the higher values of π, income will agree more with the multiple indicator index. No conclusions are to be drawn from this last statement though. These differences are too small to be of any statistical significance.

π	H	$T(x_1,x_2)$	$T(x_1,\psi)$	$T(x_2,\psi)$
10%	91	47	49	58
14%	130	70	81	86
20%	182	103	132	129
25%	227	138	175	169

Table 4 : Further results based on the multiple indicator index

As a final summary of our results, we compute the covariance matrix C between X (first three lines) and the multiple indicator index ψ.

$$C = \begin{bmatrix} 1 & & & \\ -0.3005 & 1 & & \\ 0.2099 & -0.2271 & 1 & \\ 0.5270 & -0.5700 & 0.3980 & 1 \end{bmatrix}$$

As can be seen from the last line, the multiple indicator index correlates highest with all three variables in X. Note that the correlation between ψ and X is nothing else than the vector β, since $E(X'\psi)=X'X(X'X)^{-1}\beta$ and X'X is the sample estimate of Σ. The high degree of association between the multiple indicator index and X provides an explanation as to why the classification matrices between ψ and, income and consumption respectively, offered greater scope for agreement than the analysis based on x_1 and x_2. In this sense the multiple indicator approach can be seen as a middle of the road solution between the use of income versus consumption , and vice versa.

5 Discussion

The multiple indicator approach illustrated above, can be used as an intermediary solution between the exclusive use of one indicator at the expense of others; let us consider some further points related to the use of the method.

Normative judgements have not been present in our discussion so far. And indeed, looking at our index ψ, we may be tempted to argue that it is a welfare index, based on a linear and additively separable utility function. This is certainly not the case. What ψ does is to summarize the available information about the unobservable y, this latter variable is assumed to be

pertinent to welfare analysis. All that our approach is aiming at is to draw on information from several indicators, as an alternative to using a unique noisy indicator variable. Note also that indicators such as race, sex, and age of the household head could be used as indicators of economic status. However, such household characteristics are not choice variables in a utility function in the way that bread and meat are.

The problem of constructing an index ψ from a vector of variables X, is not a new problem in multivariate statistics. As such we have suggested to work in a factor analysis set-up (the model (3.7)). An alternative, perhaps better known technique to economists, is that of *principal component analysis* (PCA). One way to obtain the PCA model would consist in omitting the disturbance term in the system relating X to y. This would then enable us to express y as a linear combination of the x's:

$$y = \delta X$$

where δ is defined as the eigen-vector corresponding to the largest eigen value of Σ [3]. In our case it is not y but E(y|X) which is constructed as a linear function of X. Thus our conclusions are about expected poverty, meaning that if two families are observationally equivalent (i.e. their observed X vectors are identical), we expect them to be equally off, however due to some unobserved factors (which the disturbance term is introduced to account for), one may be better off than the other. It is ultimately up to the researcher to decide whether PCA or factor analysis is the more adequate framework to use in the context of our problem. We have opted for factor analysis, because we feel that omitting the disturbance term would be way too unrealistic. Assuming that income, consumption, and employment status, are entirely explained by a common unobservable y excluding disturbance terms, may be too restrictive a set-up.

Throughout our presentation, we have assumed that the poverty line is to be set on a relative basis, so that the ranking of individuals would be informative of their economic status. We have not discussed the possibility of setting the poverty line on an absolute basis. Nothing in practice prevents us from doing so. We could fix a threshold ψ^*, and separate out families according to their endowments, between those who fall short of ψ^*, and those who do exceed the poverty line. The problem of adopting an absolute definition of poverty in the context of the multiple indicator approach, resides

in the fact that the quantitative position of families in the y space is not invariant to the distributional assumptions we are willing to make regarding h(y) (the distribution of y). In this sense, the multiple indicator approach offers a prediction of family ranks, rather than the absolute distances between observations, in the y space. Since y is not observed, any specification of h(y) is to be chosen as a matter of convenience. On the other hand, if progress can be made in economic theory, in order for us to have available prior specifications about the distribution of permanent income, we can make some progress towards the quantitative ranking of families; an approach more in line with the absolutist concepts of poverty. One way though to justify the normality assumption of y, would be to appeal to some argument based on the central limit theorem. If, for instance, y is the sum of many random variables (say ability, health, human capital, wealth...), we could take the view that the resulting variable is normally distributed in the population.

We have been rather vague about our definition of y, and it is often the case that investigators in the social sciences give names to the unobservable. The idea underlying this practice is the belief that the latent variables are existing well defined variables, and that the problem lies in their measurement. In practice however, whether we choose to define y as poverty, human capital, or permanent income, it makes no difference to our quantitative analysis. The unobservable y contains nothing more than the specification of the correlations between the indicators.

In the estimation of structural parameters, criteria such as consistency and efficiency will often guide the practician in his choice of estimation technique. The estimators of β and Ω we have used are derived from sample moments. In just-identified models, the *method of moments* estimator is consistent as well as *best asymptotic normal*. The method of moments estimator also presents the advantage of being distribution-free. Thus, the estimators of β and Ω will be consistent regardless of the exact form of the population distribution of X.

We now turn to the statistical properties of the multiple indicator index ψ. Firstly we note that it is a linear function of the indicators. More importantly, when the interdependence of the indicators is fully accounted for by their common dependence on y (i.e. when the axiom of conditional independence holds), the distribution h(y|X) exhausts all the information existing in the

sample about y. Since X and y are both random variables, we can decompose their joint density as follows:

$$f(X,y) = h(y \mid X) \cdot g(X)$$

Thus, all the information available to us about y, is contained in the conditional distribution h(y|X). Any index, or predictor of y, must therefore be based on h(y|X). This result is due to Bartholomew (1984), where he shows that for any g(X|y) member of the exponential family of distributions, there exist sufficient summary statistics which are linear functions of the indicators. This result is much less restrictive than one would initially think. Assume X contains p variables. The exact condition for sufficiency only requires that p-1 of the x's have distributions belonging to the exponential family. Furthermore, these distributions need not be the same. $x_1|y$ could follow a gamma distribution, while $x_2|y$ could be normally distributed etc. Note also that the exponential family is broad enough to cover many of the commonly used distributions in empirical work, such as the normal, Poisson, and gamma distributions.

As an alternative to working with the multiple indicator index ψ, one could use a weighted index of the indicators, by for example assigning equal weights to the x's:

$$\xi = (x_1 + x_2 + ... + x_p) / p$$

The weakness of an index such as ξ, based on arbitrary weights, is that generally it will not achieve the optimal reduction of the dimensionality of the problem (from p to a unique dimension) in a way that no information about y is lost. Only a sufficient statistic will achieve this optimal reduction of the dimension of the data.

Finally, when should we be content with using a single variable x_i, rather than working with the multiple indicator index? Assume we have estimated the factor analysis model (3.7), and we estimate β_i not to be significantly different from unity, and ω_{ii} not to be significantly different from zero. Under such circumstances, we could argue that x_i perfectly correlates with y, and thus that it is an ideal proxy for the unobservable.

6 Conclusions

The multiple indicator framework we have suggested is well suited to deal with a situation when a variable y pertinent to welfare analysis, is subject to measurement error. The index of y that we have proposed, is a linear function of the indicators X, and is defined as the regression function of y conditional upon X.

The multiple indicator index we have constructed from our data, offered greater scope for agreement concerning the ranking of families with both consumption and income, than in the case where families were ranked using the latter two indicators. In this sense, the proposed method in this work can be seen as a compromise with respect to those who prefer to use consumption at the expense of income, and those who favour to work with income instead of the former.

Footnotes

(1) Under some distributional assumptions, third and higher order moments may contain further information that can be used to identify the model. We therefore note that it is not always the case that with only two indicators the model is under-identified.

(2) Alternative expositions of the identification problem can be found in Goldberger(1972), and in Greene(1991) pp.531-35. The texts of Everitt (1984) and Bartholomew(1987) give a more detailed discussion of the question.

(3) See Morrison (1978) pp 267-75 .

References

Abul Naga R. (1994): "Poverty, Intergenerational Mobility, and the Role of Imperfect Information: an Inquiry with Reference to the Panel Study of Income Dynamics", Mimeo, LSE.

Anand S. and C. Harris (1991): "Food and Standard of Living: An Analysis Based on Sri Lankan Data", in Dreze J. and A. K. Sen (eds.): *The Political Economy of Hunger*, Oxford, Clarendon Press.

Anderson T.W. (1984) : *Multivariate Analysis* , New York, John Wiley.

Bartholomew D.J. (1981): " Posterior Analysis of the Factor Model", *British Journal of Mathematical and Statistical Psychology* vol. 34, 93-99.

Bartholomew D.J. (1984): " The Foundations of Factor Analysis", *Biometrika* vol. 71, 221-32.

Bartholomew D.J. (1985): " Foundations of Factor Analysis: Some Practical Implications", *British Journal of Mathematical and Statistical Psychology* vol. 38, 1-10.

Bartholomew D.J. (1987): *Latent Variable Models and Factor Analysis,* New York, Oxford University Press.

Bollen K. (1990) : *Structural Equations with Latent Variables*, New-York, John Wiley.

Chaudhuri S. and M. Ravallion (1992): "How Well Do Static Indicators Identify the Poor? ", mimeo, World Bank.

Deaton A. (1992): *Understanding Consumption*, Oxford, Clarendon Press.

Dreze J., P. Lanjouw, and N. Stern (1992) : "Economic Mobility and Agricultural Labour in Rural India: a Case Study", D.E.P. discussion paper no. 35, Sticerd, London School of Economics.

Everitt B.S. (1984) : *An Introduction to Latent Variable Models,* London, Chapman and Hall.

Friedman M. (1957): *A Theory of the Consumption Function*, Princeton, Princeton University Press.

Friedman M. and S. Kuznets (1945): *Income From Independent Professional Practice*, New York, National Bureau of Economic Research.

Glewwe P. (1990) : "Efficient Allocation of Transfers to the Poor", LSMS working paper no. 70, World Bank.

Glewwe P. and J. Van Der Gaag (1990) : "Identifying the Poor in Developing Countries: Do Different Definitions Matter?", *World Development* vol. 18, 803-814.

Goldberger A. (1972) :" Structural Equations Methods in the Social Sciences", *Econometrica* vol. 40, 979-1001.

Greene W.(1991): *Econometric Analysis*, New York, MacMillan.

Lillard L. and R. Willis (1978) : "Dynamic Aspects of Earnings Mobility" , *Econometrica* vol. 46, 985-1012.

Morrison D. F. (1978): *Multivariate Statistical Methods*, Auckland, McGraw-Hill.

Mood M. , F. Graybill, and D. Boes (1974) : *Introduction to the Theory of Statistics*, Auckland, McGraw-Hill.

Ravallion M. (1988) : "Expected Poverty Under Risk Induced Welfare Variability", *Economic Journal* vol.98, 1171-1182.

Sawhill I. (1988) : "Poverty in the U.S. , Why is it So Persistent?", *Journal of Economic Literature* vol. 26, 1073-1119.

Van Praag B., A Hagenaars, and W. Van Eck(1983): "The Influence of Classification and Observation Errors on the Measurement of Income Inequality", *Econometrica* vol. 51, 1093-1108.

Weiss Y. (1985): "The Determination of Life Cycle Earnings: a Survey" in O. Ashenfelter and R. Layard eds. : *A Handbook of Labor Economics*, Amsterdam, North Holland.