Proxy Means Test Index for Targeting Social Programs: Two Methodologies and Empirical Evidence

-Introduction. -I. I. The methods to be compared. II. Results. III. Using the poverty line to identify the poor population. -Conclusions. -Bibliography.

Primera revisión recibida agosto de 2001; versión final aceptada abril de 2002 (Eds.).

Introduction

Proxy means test indexes have successfully been used to measure household welfare (Grosh and Baker (1995)). The equations for the indicators are constructed using variables that are closely related to welfare, such as the housing dwelling's characteristics and the characteristics of the individual dwellers.

Different statistical techniques have been used to construct these indicators. For example, Grosh and Baker (1995) suggest calculating a welfare predictor using a least squares regression analysis of income (or preferably consumption) on the relevant characteristics. Others such as Castaño *et. al.* (1994) have used a qualitative principal components analysis to derive a composite index that predicts a family's living conditions.

While the economic interpretation of the first as a welfare proxy is direct by construction, the interpretation of the second indicator is not as obvious and depends on the theory of consumption's analytical framework. Similarly, compared to the first indicator, the predictive capacities of the second indicator are unknown.

The objectives of this document are to compare the predictive capacity of the two procedures for selecting social program beneficiaries and to provide empirical

evidence on the economic interpretation of the qualitative indicator discussed in Vélez, Castaño and Deutsch (1998). Based on data from the 1997 Survey of Living Conditions, we find that the variables calculated using the qualitative principal components method appear to be oriented in the same direction as the indicator explaining welfare obtained using the MORALS regression algorithm. Therefore, we find a close relationship between the indicator based on qualitative principal components and the one based on MORALS. Lastly, using the poverty line to classify the poor and non-poor populations, we find that the indicator based on principal components is as good at distinguishing these populations as the one constructed using the MORALS regression. The main conclusion is that in the absence of reliable income and consumption measures the use of the qualitative principal components method is appropriate for constructing a proxy means test.

The document is organized as follows: section 2 briefly describes the methods compared; section 3 presents the results obtained using data from the 1997 Colombia Survey of Living Conditions; the comparison of predictive capacity is in section 4; lastly, the conclusions are presented.

I. The Methods to be Compared

The two alternative procedures for constructing a proxy means test index are briefly described below.

A. The proxy means test based on a least squares regression predictor

This procedure is based on the assumption that income or consumption are appropriate *measurements* for the welfare of individuals and that variables related to housing and individual's conditions are highly correlated with these measurements. Grosh and Baker (1995) suggest constructing the proxy means test predictor using a regression in which the dependent variable is the household's per capita income or household's consumption (preferably consumption), and the explanatory variables comprise a set of welfare-related, individual and household measures. In general, the equation to obtain this predictor is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i \qquad (1)$$

where y_i is the logarithm of income (or consumption), and x_{ji} are the welfare-related characteristics of the dwelling and its residents, β_j are the least square parameters estimates, and ε_i is the random error term.

The predictor $\hat{y}_i$ obtained using this equation is an approximate measurement of household welfare and, using the poverty line, the potential beneficiaries of a social program can be chosen.

Because the explanatory variables used here are qualitative (non-dichotomous, in general) and quantitative, the regression suggested by Grosh and Baker (1995) cannot be applied to them. To solve this problem, an alternative procedure of multiple optimal regression, called MORALS, was used. This is a procedure that assigns quantitative values to the categories of categorical variables in order to maximize the regression's *coefficient of determination*. Young (1981) and Young, de Leeuw and Takane (1976) present a discussion of the algorithm. Once the explanatory variables have been quantified, they can be entered into a least squares regression. An application of this technique can be found in Castaño (1999c). If we assume that x^*_{ij} are the quantified variables, then the equation to get the welfare predictor is simply:

$$y_i = \beta_0 + \beta_1 x^*_{1i} + \beta_2 x^*_{2i} + \ldots + \beta_p x^*_{pi} + \varepsilon_i \qquad (2)$$

In other words, a least squares regression of y_i on the x^*_{ij} quantified variables is run.

B. The proxy means test index based qualitative principal components

An alternative method for constructing the indicator involves using a Principal Components procedure to derive a weighted index of the variables. As in the case above, the presence of qualitative and quantitative variables prohibits the use of this analysis. As before, we can quantify the variables. The procedure is described in Young (1981), Kuhfeld, Sarle, and Young, (1985), Saporta, (1983), Young, Takane, and de Leeuw, J. (1978, 1985), Van de Geer, (1993) and some applications are found in Castaño et al (1994), Sarmiento et al (1996), Castaño, Correa and Salazar (1998), Castaño and Valencia (1999a), and Castaño (1999b).

The indicator is obtained as the first principal component of a system of quantified qualitative variables. In this case, the quantification algorithm assigns numerical values to the categories of the variables in such a way that maximizes the variance of the first principal component. If we assume that z^*_{ij} are the quantification of the x_{ij} characteristics, then the indicator is the first principal component of the z^*_{1j}, z^*_{2j}, ..., z^*_{pj}, variables system, which takes the following form:

$$CP1_i = \alpha_1 z^*_{1i} + \alpha_2 z^*_{2i} + \ldots + \alpha_p z^*_{pi} \qquad (3)$$

The variable $CP1_i$ contains maximum information on the z^*_{1j}, z^*_{2j}, ..., z^*_{pj} system. In this case, $CP1_i$ is not observed (as in the case of equation (2)); instead, it is *constructed* once the α_j coefficients have been estimated. Clearly, in this case there is no dependent variable to help quantify and the resulting question is whether there is any relationship between the indicators derived through these two different methodologies. For example, it is important to know the type and strength of the relationship between the two indicators, the relationship between the principal components method's quantifications (that does not take consumption into account) and consumption (or some function for consumption), and the predictive capabilities of this indicator in comparison to the indicator constructed using regression.

To answer this question, both indicators were constructed using data from the 1997 Colombia Survey of Living Conditions. Some of the results are presented below.

II. Results

These results are based on data collected in the 1997 Survey of Living Conditions. A set of variables thought to be correlated with household per capita consumption, was chosen a priori. These variables represent different aspects of welfare conditions, such as quality of the dwelling, access to public services, ownership of durable goods, demographic aspects, schooling, social security (health), and employment and work conditions.

From a large set of variables initially selected, the following remained in the analysis due to their capacity to predict consumption:

a. On housing conditions and location.

- Location of dwelling given by the dwelling's electric power stratum (ESTRAT).
- Quality of the floors in the dwelling (MATPIS).
- Cooking fuel (CONQCOC).
- Number of toilet facilities (NUMSAN).
- Shower (DUCHA).

b. On services in the dwelling:

- Installation of a telephone line (TELEF).
- Sewage disposal (EXCRE).
- Garbage disposal (BASUR).

c. On ownership of durable goods:

- Refrigerator (NEVERA).
- Video recorder (BETA-VHS).
- Car (AUTO).
- Washing machine (LAVAD)
- Cable television (TVCABLE).
- Heater (CALENT).
- Motorbike (MOTO).
- Air conditioner (AIRE).
- Vacuum cleaner (ASPIR).
- Color television (TVCOLOR).

d. On demographic factors:

- Crowding: number of rooms per person (HACIN).
- Number of children six and under (NPM6).
- Sex of head of household (JEFSEX).
- Age of head of household (EDADJ).

e. On schooling:

- Educational attainment of head of household (ESCJEF).
- Educational attainment of spouse of head of household (ESCJEF2).
- Attendance at official primary school.
- Attendance at official secondary school.

f. On health, job skills, and income recipients:

- Proportion of earners (PRPERCE).
- Proportion of insured individuals in the household (PRPERSS).
- Occupational Status (POSOC).

For the construction of the regression-based indicator, the MORALS algorithm in the SAS statistical package was used via the TRANSREG procedure. To avoid heteroskedasticity and normalize the distribution of income, the natural logarithm of household per capital consumption was used as a dependent variable. This makes the categories of the qualitative variable quantifiable with respect to this variable and not directly to consumption.

To construct the indicator based on qualitative principal components, the MTV algorithm of the SAS statistical program was used using the PRINQUAL procedure. Note that this procedure does not use the consumption variable and that the estimates need not explain welfare in terms of consumption.

Some of the most important results that allow us to compare the two procedures are presented below.

A. Relationship of the quantifications of the two methods with respect to the household per capita consumption logarithm.

To begin to understand the possible relationship between the indicators, it is important to measure the associations between the per capita consumption logarithm and each of the variables quantified using the two methods.

Table 1 presents the correlations of the variables quantified using both methods, with the household per capita income logarithm. The column headed REGRESSION contains the correlations of the variables quantified using the MORALS algorithm and the column headed COMPPRIN contains the correlations of the variables quantified using the qualitative principal components algorithm.

Table 1. Correlations of the Quantified Variables and the log (Per Capita Consumption)

Variable	Regression	Comprin
ESCJEF	0.57826	0.58761
ESCJEF2	0.45100	0.45023
ESTRAT	0.58456	0.59781
MATPIS	0.49118	0.50864
POSOC	0.45978	0.50108
NUMSAN	0.48859	0.49449
TELEF	0.41043	0.48902
HACIN	0.44654	0.47057
PRIMOF	0.32387	0.31910
SECOF	0.26761	0.27574
PRPERCE	0.32247	0.31644
PRPERSS	0.28750	0.28521
CONQCOC	0.18546	0.23131
NPM6	0.21551	0.22161
EDADJ	0.11177	0.07050

All the correlations are positive and, in general, their magnitude is very similar. It is evident that in many cases the principal components method, which

does not use consumption information, produces simple correlations that are greater than those produced by the regression method[1] which does uses the consumption variable. In both cases, the variable with the weakest association is age of head of household (EDADJ). From the discussion above, we conclude that the principal components method, without using consumption, appears to approximate the logarithm of that measurement of welfare.

B. Description of the indicators and their correlation

Table 2 shows descriptive statistics for the two indicators. Hereafter, the regression-based indicator will be referred to as IREGR and the principal components-based indicator as ICOMP

Note the strong association between the two indicators. This confirms the previous finding and indicates that if the regression-based indicator together with the poverty line can be used to identify the potential beneficiary population, perhaps

Table 2. Descriptive Statistics

Variable	Mean	STD	Minimum	Maximum
IREGR	11.595348	0.897507	9.181362	14.719929
ICOMP	0.0000000	2.7264943	-6.4064199	9.5812052
Correlation between the indicators: **0.93834**				

the components-based indicator could also be used. But the problem is that this indicator does not have the same units of measure as the regression-based indicator. One solution to this problem is to replace the origin and scale with that of the regression indicator, using the following linear transformation:

$$TICOMP = Mean(IREGR)+STD(IREGR)*(ICOMP-Mean(ICOMP)))/ STD(ICOMP) \quad (4)$$

1 The fact that the MORALS method provides some transformed variables with smaller simple correlations than the qualitative principal components method is interesting to analyze. Remember that the objective of MORALS is to obtain quantifications in such a way that maximizes the coefficient of determination R2. Therefore, another procedure is unlikely to provide quantifications that when regressed on the dependent variable, achieve a better fit than that obtained with MORALS. In fact, if we get the R2 of the regression using the PRINQUAL quantifications, we get an adjustment of 0.627 versus 0.677 with MORALS. This means that getting higher simple correlations does not necessarily imply a better "overall" explanation of welfare.

where TICOMP is the linear transformation over ICOMP, Mean(X) and STD(X) are respectively the mean and the standard deviation of the X variable. Note that this indicator has the same mean and variance as IREGR, as shown in Table 3.

With the transformed TICOMP indicator we can use the poverty line, since it is in the same units as IREGR.

Table 3. Descriptive Statistics of IREG and TICOMP

Variable	Mean	STD	Minimum	Maximum
IREGR	11.5953479	0.8975066	9.1813620	14.7199289
TICOMP	11.5953479	0.8975066	9.4864845	14.7492862

III. Using the Poverty Line to Identify the Poor Population

Note that the above indicators are associated with the welfare logarithm (represented as the household per capita consumption logarithm). To obtain a welfare logarithm, and thus the distribution of welfare in the population, we must retransform the predictor in the previous equation (IREGR) and perform the same transformation on TICOMP.

A bias corrected transformation for IREGR, denoted by RIREGR, is (Castaño, 1997),

$$\text{RIREGR} = e^{\text{IREGR} + 0.5\hat{\sigma}2} \qquad (5)$$

where e is the exponential function and $\hat{\sigma}^2$ is the estimated error term variance of the least squares regression of the natural logarithm of per capita consumption on the quantified variables.

The analogous retransformation for TICOMP, denoted as RTICOMP would be:

$$\text{RTICOMP}=e^{\text{TICOMP} + 0.5\hat{\sigma}2} \qquad (6)$$

This last expression would be a proxy of welfare based on household per capita consumption. What is its predictive capacity compared with RIREGR?

The poverty line for 1997 was $107099.71. Using this cutoff to determine the percentage of poor population, we get the following percentages for each indicator:

Percentage using the poverty line: 53.0%

Percentage using the regression indicator (RIREGR): 45.6%

This and previous results indicate a strong association between the approximate welfare by household per capita consumption and the indicator generated by the qualitative principal components technique.

Percentage using the principal components indicator (RTICOMP): 49.8%

The above results show that the RTICOMP indicator seems to work better than the RIREGR, since it identifies a percentage of poor population closer to that obtained when using the poverty line (PL).

However, we do not know which were the poor households identified by the different methods of poverty line, RIREGR, and RTICOMP. To clarify this doubt, the classification tables for each method appear below.

The tables above indicate that the poverty line method classifies 2719 households as poor. Of those, 2099 are also identified as poor by IREGR (Table 4) and 2136 are identified as poor by TICOMP (Table 5).

Table 4. Classification using RIREGR

		Poor	Not poor	Total
	Poor	2099	620	2719
		77.20	22.80	
PL	Not poor	335	1858	2193
		15.28	84.72	
	Total	2434	2478	4912

Table 5. Classification using RTICOMP

		Poor	Not poor	Total
	Poor	2136	583	2719
		78.56	21.44	
PL	Not poor	489	1704	2193
		22.30	77.70	
	Total	2625	2287	4912

Conclusions

Using the data from the 1997 Colombia Quality of Life Survey, we have provided empirical evidence that confirms a close relationship between a proxy means index generated using qualitative principal components and the proxy means index based on a regression of consumption as an indicator of welfare. The fundamental conclusion is that the use of qualitative principal components is

suitable to construct proxy means tests, in the absence of sources of information that include reliable measurements for income and consumption.

Bibliography

CASTAÑO, E. and MORENO H. (1994) "Selección y cuantificación de las variables del Sistema de Selección de Beneficiarios, Sisben", *Planeación & Desarrollo*, Vol. XXV, Julio de 1994.

CASTAÑO, E. (1997) "Estimación de la media en poblaciones asimétricas", *Revista Colombiana de Estadística, 34.*

CASTAÑO, E., CORREA, C. and SALAZAR B. (1998), *La construcción de un indicador de calidad de vida para la ciudad de Medellín*, Santafé de Bogotá, DNP, Misión Social. Mec.

CASTAÑO, E. and VALENCIA, L. (1999a) "Metodología Estadística para determinación de los estratos en la ciudad de Medellín", *Lecturas de Economía, 50.*

CASTAÑO, E. (1999b), "Diseño del sistema de identificación de beneficiarios de programas sociales", Quito, Consejo Nacional de Modernización del Estado, Conam. Mec.

CASTAÑO, E. (1999c), *Diseño de un indicador para la focalización de programas sociales en Argentina*, Buenos Aires, Secretaría de Desarrollo Social, Siempro.

GROSH, M., and BAKER, J. (1995), "Proxy Means Tests for Targeting Social Programs: Simulations and Speculation", *Living Standard Measurement Study Working Paper* No. 118. World Bank.

KUHFELD, W.F., SARLE, W.S. and YOUNG, F.W. (1985) "Methods for Generating Model Estimates in the PRINQUAL Macro", SAS Users Group International Conference Proceedings: Sugi 10, Cary, NC:SAS Institute, 962-971.

SAPORTA, G. (1983), "Multidimensional data analysis and quantification of categorical variables", *New Trends in Data Analysis and Applications*, J. Janssen, J.F. Marcotorchino, J.M. Proth Eds., Elsevier Science Publishers B.V., North-Holland.

SARMIENTO, A., RAMÍREZ, C., MOLINA, C. and CASTAÑO E. (1996), *Índice de condiciones de vida*, Santafé de Bogotá, DNP-Misión Social. Mec.

SAS/STAT User Guide (1990), Volume 2, Version 6, Fourth edition.

VAN DE GEER, J.P. (1993), *Multivariate Analysis of Categorical Data*, London, Sage Publications.

VÉLEZ, C.E., CASTAÑO, E., and DEUTSCH, R. (1998) "An Economic Interpretation of Colombia´s SISBEN: A composite Welfare Index Derived from Optimal Scaling Algorithm", First Workshop of LACEA/IDB/World Bank Inequality and Poverty Network.

YOUNG, F.W. (1975), "Methods for Describing Ordinal Data with Cardinal Models", *Journal of Mathematical Psychology*, 12, 416-436.

YOUNG, F.W. (1981), "Quantitative Analysis of Qualitative Data", *Psychometrika*, 46, 357-388.

YOUNG, F.W., TAKANE, Y. and de LEEUW, J. (1978), "The Principal Components of Mixed Measurement Level Multivariate Data: An Alternating Least Squares Method with Optimal Scaling Features", *Psychometrika*, 43, 279-281.

YOUNG, F.W., TAKANE, Y. and de LEEUW, J. (1985), *PROC PRINQUAL- Preliminary Specifications*, [unpublished manuscript], Chapel Hill, The University of North Carolina Psychometric Laboratory.