FINDING THE POOR

BY ROBERT FERBER
University of Illinois

AND

PHILIP MUSGROVE
Brookings Institution

As a basis for judging how public policy affects the poor, this article explores how "poor" families may be defined and how well such families can be distinguished from other families in the less developed countries. This is done by seeking proxies for poverty which are relatively easy to measure, accurate in discriminating between the poor and the non-poor, and relevant to public policy. To this end, a highly parsimonious model is developed, based on truncation and regression procedures, using only family size and number of wage earners in addition to either income or an education-age combination. Application of this model to data from household surveys in three major cities of Latin America shows that the model is highly effective in pinpointing poverty households, although the pattern of errors is not random, the most frequent type of error being to classify poverty households as non-poor.

Especially significant is that the model is nearly as effective for discriminating poverty households from others when financial variables are excluded as when they are included. This would suggest that a good deal of flexibility exists in deciding what variables to include in future studies of this type. The results also suggest that even better results should be possible if more complete information is obtained on the employment status of the different members of the household and on the contribution of each to household income. Ideally, the data collection and model development should proceed in an iterative manner since there are numerous possible variables as well as alternative model formulations.

1. INTRODUCTION

This paper reports the results of a pilot study undertaken with data from three household surveys in Latin America to ascertain how well the poor can be identified by characteristics obtained readily through household surveys. The focus is on a methodology for doing so which, while tested on three cities in South America, also seems to have applicability to other countries, including the United States.

The reason for studying this problem stems from the fact that the poorest families in virtually all countries remain in dire poverty even though significant improvements may be achieved in income per head or other measurements of economic progress. Since the poor do not share automatically in economic development, it becomes urgent to judge public policy by how well it reaches poor families as well as by its effects. This requires a means of identifying "the poor", to define which families are in poverty and how they are distinguished from the non-poor. To do so by going to "poor" neighborhoods and selecting households with low incomes is not as effective as it may seem, especially in the less developed countries where household income is a more nebulous concept (partly because much of it is in kind). Also, the "poor" do not necessarily live in "poor" neighborhoods.

In this setting, "finding the poor" takes on a more significant meaning. It means finding proxies or indicators for poverty, characteristics which are (1) relatively easy to establish or measure, (2) accurate in discriminating between the poor and the non-poor, and (3) relevant to the design or evaluation of public policies. An ideal proxy will divide households into groups that are easily identified and can be reached by public action, and such that there are large differences in welfare among groups but only small differences within groups.

The remainder of this paper is divided into six parts. The next part presents a brief discussion of the nature of the data and of the statistic used to measure poverty and its rationale. Section 3 then summarizes some exploratory investigations of the relationship of this measure to available characteristics, which then leads into the basic model. The empirical results obtained with the model are summarized in Part 5, with validation tests of the model covered in Part 6. The concluding section summarizes the principal results and suggests directions for further work.

2. Background Information

After reviewing some general approaches to the concept of poverty and comparison of alternative measures, the definition of poverty selected for this study is *per capita* consumption expenditures (C/N).[1] In comparison to income or some other measure of well being, it was clear from the data that consumption expenditures *per capita* yielded a much more stable and meaningful measure than any of the others. Moreover, all indications pointed to a much higher level of reliability for the consumption data than for the income data in these household surveys, which was all the more to be expected because of the very great detail in which consumption expenditures were obtained in these surveys.

Use of total expenditures rather than basing poverty on some component of consumption seems desirable partly because of the highly variable nature of these components among poverty families with different characteristics, and partly because of the difficulty of applying the latter approach to cities in different countries under different cultural conditions.

Placing this measure on a *per capita* basis serves to provide a more realistic indication of the level of living of the family. No attempt was made in this exploratory study to adjust these *per capita* computations for the different ages of the family members, since indicators from other work with these data suggest that this adjustment would have little effect on the location of individual families in the income distribution.[2]

Consideration was also given to defining poverty on the basis of characteristics of the family and of the housing unit, such as condition of the unit and

[1]Musgrove, Philip and Ferber, Robert, *Identifying the Urban Poor*, to be published in *Latin American Research Review*.

[2]Research by Aquiles Arellano ("Hacia una Canasta de Consumo Minimo", Working Paper, Universidad de Chile, Santiago, August 1975) finds that the cost of a subsistence diet is very nearly the same for adults, adolescents and young children, being appreciably lower for infants. When subsistence expenditure, both for food alone and for all spending together, is related to family size, the elasticity is about 0.9; since large families consist more of children, this is further evidence that there are no great differences between adults and children.

employment status of the family members. However, such an approach would have necessitated a number of rather arbitrary assumptions, and the relationship of these variables to poverty was very unclear, at least for these cities.

The Data

The data used were collected as part of a program of household budget surveys undertaken by ECIEL (Programa de Estudios Conjuntos sobre Integración Económica Latinoamericana), a consortium of research institutes in 17 cities of nine countries of Latin America. The surveys in Bogota and Medellin, Colombia, were undertaken in 1967–68 and the one in Lima, Peru, in 1968–69. The surveys are described in detail in publications of the institutes which collected the data and shared in their analysis.[3] Suffice it to say here that approximately 800 families were interviewed in each of the Colombian cities and 1,357 families in Lima.

Data were collected in these surveys by a standardized questionnaire on a wide number of family characteristics, including family composition, employment status and occupation of each of the family members, mobility, consumption expenditures in considerable detail, and income by type though not income earned by each wage earner separately. Special attention was given to inclusion of non-monetary income and consumption since such items were known to constitute a large portion of the resources of many of these families. The scope of the study is perhaps best indicated by the fact that approximately 1,000 variables were coded for these families, of which about 550 relate to consumption expenditures.[4]

For the purposes of this study it was decided to use a simple dichotomous criterion for classifying poverty families, namely, those families that were in the lower 40% of the distribution by the poverty measure. This criterion was selected because to classify families on the basis of minimum needs for subsistence was hardly feasible for those cities, considering the sparse data available for this purpose, so that some cutoff point for the distribution of actual expenditures *per capita* seemed much more meaningful. Since previous indications were that at least one-third of the families in these cities would be classified as being in poverty by almost any reasonable measure, and since it was desired not

[3]See Rafael Prieto Duran, *Estructura del Gasto y Distribución del Ingreso Familiar en Cuatro Cuidades Colombianas, 1967–68* (Bogota: Universidad de los Andes, 1971), and Adolfo Figueroa Arevalo, *Estructura del Consumo y Distribución de Ingresos en Lima Metropolitano, 1968–69* (Lima: Pontificia Universidad Catolica del Peru, 1974).

[4]The ECIEL household data (collected in 17 cities in nine countries, and analyzed thus far for 11 cities in six countries) have already been used to study a number of features of the urban income distribution. See Philip Musgrove, *Income and Spending of Urban Families in Latin America, The ECIEL Consumption Study* (Washington: The Brookings Institution, 1978). In addition, much use has been made of the data to examine how spending on different categories varies with income (or total consumption) and with a variety of household characteristics. See Musgrove, *Income and Spending*; the studies by Prieto (Colombia) and Figueroa (Peru) cited earlier; Howard J. Howe, "Estimation of the Linear and Quadratic Expenditure Systems: A Cross-Section Case for Colombia" (Ph.D. thesis, University of Pennsylvania, 1974); and Howard J. Howe and Philip Musgrove, "An Analysis of ECIEL Household Budget Data for Bogota, Caracas, Guayaquil and Lima", Chapter 7 of Constantino Lluch, Ross Williams and Alan Powell, *Patterns in Household Demand and Saving* (Oxford University Press, 1977).

to omit families that might be in poverty, the use of the fortieth percentile as a cutoff point seemed quite reasonable.[5]

3. Preliminary Eplorations

As a basis for devising a model for identifying the poor, a number of exploratory analyses were carried out relating the measure of poverty, as measured by consumption *per capita*, to a host of other demographic and socioeconomic variables collected in the surveys. The pertinent results that led to the development of the model presented in the following section may be summarized briefly as follows:[6]

1. In terms of income, poverty families are characterized by one or two wage earners each having a low labor income and with little or no income from transfers, capital or other sources.

2. The higher is the dependency burden (the ratio of the total number of family members to employed family members), the more likely the family is to be poor.

3. Hardly any poverty families were found to be located in the high stratum of the neighborhood stratification design used in these surveys. In each of the three cities, neighborhoods had been stratified on the basis of *a priori* information as "high", "middle" and "low" on the basis of various socioeconomic criteria. The frequency of poverty of families varied from 2 percent of those in the high stratum to approximately 50 percent of those in the low stratum.

4. A lack of water or electricity in a dwelling is an almost sure sign of poverty in Bogota, a very likely sign of poverty in Medellin and less so in Lima. Thus, of the households lacking water, all of them in Bogota fell in the poverty classification, as did 86 percent of those in Medellin but only 62 percent of those in Lima.

5. Other dwelling unit characteristics, such as tenancy and type of construction of the unit, do not show much relation to poverty. There was some tendency, however, for the likelihood of a family to be in poverty to increase as the number of family members per sleeping room increased, once this density figure exceeded one.

6. Age of head has a U-shaped relationship with poverty, the likelihood of a family being in poverty being highest at lower ages and at higher ages. However, this effect is strongly affected by education, being much less pronounced among those heads that have more education.

7. Education is strongly and negatively correlated with poverty status. In particular, virtually no families in any of these cities with a college education were likely to be in these poverty classifications.

8. Occupation of head is also correlated negatively with poverty, in the sense that the relative frequency of poverty of families in a particular occupa-

[5]Of course, one need not use a dichotomy at all, but rather simply see how well a model can reproduce the actual distribution in terms of the poverty measure. However, for the present purposes, and also for most policy purposes, such an approach is stringent. Admittedly, however, one might care to refine the approach developed here to have a trichotomy, in order to segregate families that are in the most extreme state of poverty.

[6]These analyses are contained in Ferber and Musgrove, *op. cit.*, Parts Three and Four.

tional category tends to decline as the skill demands of that occupation increase. However, the relationship is much less strong than in the case of education.

9. Sector of employment of head shows little relationship to poverty except for a tendency for more families in poverty to have a head employed in the construction industry.

These findings lead to two conclusions about the identification of poverty by *ex ante* socioeconomic and household characteristics. One is that since these various variables interact with each other, some form of multivariate analysis is needed to ascertain how well these variables as a group serve to pinpoint poverty families. The second is that it should be easier to use such relations to pinpoint poverty if the sample is first truncated.[7] In other words, households which by simple criteria are almost certain to be poor, or certain not to be poor, are best singled out in advance, so that any multivariate analysis can be focused on those households for which classification is more difficult. The results also suggest that it may be easier to find and exclude non-poverty families than poverty families.

4. Development of a Model

Provision for Validation

Since it is clear from the foregoing results that alternative specifications and different combinations of variables may have to be tested in the search for better relationships, there is a critical need to make provision for the detection of spurious relationships that may result from sampling variations and quirks in the data.

The best-known means of dealing with this problem is to divide the data set by a statistically random process into two parts, test alternate relationships on one part and, after a "best" specification is obtained, apply the same specification to the other part. If the results from the first sample are fully valid, similar results should be obtained from the second sample (at least within the margin of sampling error); and to the extent that the second sample yields a different (worse) result, evidence of search bias is obtained.

This procedure is followed in the present study. In each city, the data set is randomly divided into two equal-size samples, namely, Sample *A* and Sample *B*. In each case, Sample *A* serves as the "test sample", on which alternative model formulations are tested and a "best" model is obtained. That model is then fitted to the Sample *B* data and the results compared with those of Sample *A*.

Truncation

From the results obtained earlier, it would seem logical to try to truncate the distribution of families by the poverty measure (consumption *per capita*) at both ends, that is, by using some variables that clearly identify families that are not in

[7]The alternate procedure of a dummy variable regression with poor–nonpoor as the dependent variable using all the observations was discarded because the truncation procedure offered the possibility of prior elimination of groups that could clearly be identified as poor, thereby using the regression procedure for the more difficult classification groups.

poverty and using other variables that clearly identify families that are in poverty. Since these are two extremes, it is best to consider each separately.

The Non-Poor. An obvious variable for singling out families not in poverty, from the foregoing results, is location in the "high" sample stratum. We denote this stratum by A_1.

It would also seem feasible from the foregoing results to proceed one step further and to select from the remaining families, A_2, a set of families which, although living in "middle" or "low" strata, are also very unlikely to be poor. For this purpose, we use information about three kinds of assets—human capital, financial wealth and physical capital. A family is classified into this group, A_{21}, if:

the head of the household has university or post-secondary education, or
the family owns a car, or
the family has a bank account (either checking or saving).

The Poor. Poverty appears to be best indicated by characteristics of the dwelling and by family composition. At this end, therefore, we truncate a subset of A_2 as being in poverty, A_{23}, if:

the proportion of adults in the household is below 20 percent (four or more children per adult), or
the dwelling lacks piped water, or
the dwelling lacks electricity, or
there are more than four people per sleeping room.

Also classified as poor are households where nobody is employed, at least one adult is looking for work and the head is not retired. None of these latter criteria is likely to exclude many families, it being extremely difficult to locate large numbers of poor households by any criterion that does not also include many non-poor families. However, these criteria are readily applied and, judging by the exploratory results, may be highly effective for the present purposes.

An overall view of the procedure is provided in Figure 1. Working with Sample A, the object is to define a subset A_{22} whose members are not readily classified *ex ante* as either poor or non-poor, but who are characterized by stable relations between C/N and a small number of non-financial characteristics. Regression analysis is then used to pinpoint which families in A_{22} are poor and which are not, and thereby to identify those ranges and combinations of variables associated with poverty.

We reduce A to A_{22} in three steps, indicated by the equation

$$A_{22} = ((A - A_1) - A_{21}) - A_{23}.$$

Here A_1 and A_{21} both represent groups which can be identified as non-poor, allowing the upper end of the distribution of C/N to be truncated. A_{23} represents households identifiable as poor, truncating the low end of the distribution. The reason for the asymmetry—removing the rich in two steps but the poor in only one step—is entirely a consequence of the structure of these samples, in particular of their *ex ante* socioeconomic stratification.

We chose not to use several other criteria which would locate a few more families in poverty along with nearly equal numbers of non-poverty families, criteria such as housing type, unemployment and migration status. Neither have

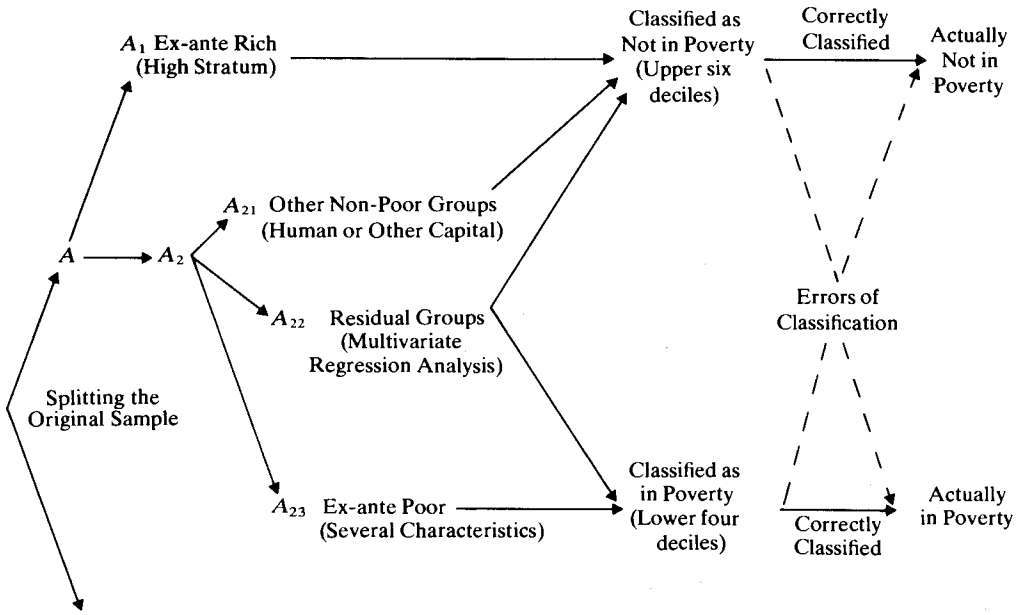

Figure 1. Model for Classifying Families as Poor or Non-Poor, and Testing the Classification Accuracy

we used education (except at the upper end), age, life cycle or dependency burden (except again at the upper end) because of the feeling that such variables are best used jointly in the multivariate analysis.

The Residual Group

The truncated set of households, A_{22}, is differentiated by poverty status using a number of alternative regression models. Regression analysis was used rather than some form of discriminant analysis partly because it was felt desirable for policy purposes to be able to see how well the models detected households in extreme poverty circumstances, such as in the first decile. Also, it was felt that some experimentation might be needed with different cutoff points for the definition of poverty, and for these purposes the regression approach was the most flexible.[8]

In formulating such models, two alternative approaches were followed, one using only nonfinancial variables and the other using a financial variable as well. The reason is that while a model with financial variables may be more successful, such variables are much more difficult (and expensive) to collect. It is of interest, therefore, to ascertain the extent to which a model using easily-obtainable nonfinancial variables is equally successful. The margin of difference, to the

[8]While logit or probit analysis might have been used also, the fact is that "poverty", as we have defined it, is a judgmental distinction on a continuum, not a natural binary variable.

extent it exists, is of crucial importance in evaluating the potential benefits of future, more expensive survey designs for collecting such financial variables.

The function with financial variables includes income of the head of the household (Y_H) and the dependency ratio, expressed as the ratio of employed adults to total members in the households (N_W/N). The early results of the exploratory stage suggest that these variables are related to the poverty measure. They do not suggest, however, whether N_W/N is best considered as a single variable or whether each term is better considered a separate component, N_W measuring the wage earner effect and N the household size effect. It is also not clear whether these separate terms should be arithmetic or logarithmic to allow for possible scale effects.

For these reasons four financial-variable functions were tested, namely:

$$(1) \qquad \log \frac{C}{N} = \beta_0 + \beta_1 \log Y_H + \beta_2 N_W + \beta_3 N$$

$$(2) \qquad \log \frac{C}{N} = \beta_0 + \beta_1 \log Y_H + \beta_2 \log N_W + \beta_3 \log N$$

$$(4) \qquad \log \frac{C}{N} = \beta_0 + \beta_1 \log Y_H + \beta_2 \frac{N_W}{N}$$

$$(4) \qquad \log \frac{C}{N} = \beta_0 + \beta_1 \log Y_H + \beta_2 \frac{N_W}{N} + \beta_3 \log N$$

If scale effects are not important, β_2 and β_3 in (2) will be unity, but with opposite signs. If the wage earner effect is of the same magnitude as the household size effect, β_2 will equal $-\beta_3$ in (1) and (2). Also, if in (2) or (4), β_1 is equal to $-\beta_3$, the household size effect can be merged with the income effect to yield *per capita* income (but, in this case, income of the head) as a determinant of *per capita* consumption.

For the function with nonfinancial variables, income is replaced by what seem to be its two principal determinants, which are easy to obtain in a survey—education and age, both of the head. The interaction of these variables could be specified in dummy variable form, but this has the disadvantage of producing sharp jumps in income from one age group to another and of requiring a large number of coefficients to deal with just a few classes of each variable. It seems preferable to introduce age as a continuous variable, and the anticipated curvature of the age-income profile can then be allowed by introducing a quadratic term as well. This approach had previously been used successfully for Bogota and Mendellin[9].

Based on the preceding results, it also seems desirable to allow the education dummy variables to interact with the age variables, producing as many different age-income profiles as there are distinct education classes. This leads to

[9]Howe, "Linear and Quadratic Expenditure Systems," *op. cit.*

the following four specifications:

$$(5)\quad \log\frac{C}{N}=\sum_{i=1}^{5}\beta_{1i}E_i+\sum_{i=1}^{5}\beta_{2i}E_iA+\sum_{i=1}^{5}\beta_{3i}E_iA^2+\beta_4N_W+\beta_5N$$

$$(6)\quad \log\frac{C}{N}=\sum_{i=1}^{5}\beta_{1i}E_i+\sum_{i=1}^{5}\beta_{2i}E_iA+\sum_{i=1}^{5}\beta_{3i}E_iA^2+\beta_4\log N_W+\beta_5\log N$$

$$(7)\quad \log\frac{C}{N}=\sum_{i=1}^{5}\beta_{1i}E_i+\sum_{i=1}^{5}\beta_{2i}E_iA+\sum_{i=1}^{5}\beta_{3i}E_iA^2+\beta_5\frac{N_W}{N}$$

$$(8)\quad \log\frac{C}{N}=\sum_{i=1}^{5}\beta_{1i}E_i+\sum_{i=1}^{5}\beta_{2i}E_iA+\sum_{i=1}^{5}\beta_{3i}E_iA^2+\beta_4\frac{N_W}{N}+\beta_5\log N$$

Five categories E_i are used for education—none, some primary, complete primary, some secondary, and complete secondary. Higher levels of education are unnecessary because they have been included in subset A_1. A constant term was initially omitted from these equations to allow the coefficients of all the dummy variables to be estimated. However, multicollinearity among the variables did not permit inclusion of education, education-age and education-age squared terms in the same function. On the basis of various tests as well as *a priori* reasoning, it was felt more important to retain the interaction terms and therefore education as a separate variable was dropped from the equations and constant terms were included, leaving

$$\beta_1+\sum_{i=1}^{5}\beta_{2i}E_iA+\sum_{i=1}^{5}\beta_{3i}E_iA^2$$

and the terms in N_W and N.

In addition to these functions, it seemed desirable to make a similar series of tests using a linear form rather than a logarithmic form. While a logarithmic form would ordinarily be considered more desirable when monetary magnitudes are involved, as in the present case, it should be remembered that the truncation procedure, if successful, will have removed the extremes of the distribution. Hence, the factors that would normally be expected to show scale effects and be responsible for curvature may not be very important. Whether this is true is not clear without empirical tests, and since Sample A is designed precisely for such testing, a set of linear functions are fitted to the data as well, corresponding to the eight logarithmic functions outlined previously.

5. Empirical Results

Sample A

As is evident from Table 1, the results obtained from applying the truncation procedure are much better at distinguishing households not in poverty than households that are in poverty. In particular, it is evident from the table that the procedure of automatically classifying all households in the high sample stratum as nonpoor (the set A_1) is extremely effective, with only 2 to 5 percent of

TABLE 1

ACCURACY OF TRUNCATION PROCEDURE, UNWEIGHTED DATA, SAMPLE A

Stratum	Households Classified in Stratum		Errors	
	Number	Percent of Total	Number	Percent of Stratum
		Bogota		
A_1 (nonpoor)	57	15	3	5
A_{21} (nonpoor)	66	17	8	12
A_{23} (poor)	72	19	16	22
A_{22} (residual)	193	49		
Total	388	100		
		Medellin		
A_1 (nonpoor)	80	21	2	3
A_{21} (nonpoor)	72	19	7	10
A_{23} (poor)	71	18	19	27
A_{22} (residual)	161	42		
Total	384	100		
		Lima		
A_1 (nonpoor)	180	28	3	2
A_{21} (nonpoor)	175	27	13	7
A_{23} (poor)	133	21	50	38
A_{22} (residual)	154	24		
Total	642	100		

the sample of households misclassified by this rule. The next step, separating out from the remainder those households that have a well-educated head or own certain assets, yields somewhat higher misclassifications but is still very satisfactory, with errors ranging between 7 and 12 percent. Thus, these two sets together seem to weed out successfully substantial proportions of the sample observations as being nonpoor, the overall error ranging from about 4 percent from Lima to about 9 percent for Bogota.

Much less successful is the truncation at the other end of the distribution. Here, the attempt to separate out from the remaining households those in poverty on the basis of housing density and lack of certain utilities yields errors of misclassification ranging from 22 percent for Bogota to 38 percent for Lima. Further investigation suggests that the housing density rule is too liberal, and that better results might be obtained if that criterion were altered or perhaps eliminated altogether.

The figures in Table 1 refer to sample sizes, and are not adjusted for different sampling ratios used, for example, to overrepresent the higher income areas. For this reason, summing the different strata will not yield accurate indications of the error rates to be expected in the population from the truncation procedure; this is done later. At this stage, focus on the unweighted data is desirable, however, since it brings out more clearly how classification error

TABLE 2

ADJUSTED VALUES OF R^2 FOR ALTERNATIVE REGRESSIONS FITTED TO STRATUM A_{22}

Independent Variables	Bogota	Medellin	Lima
1. Log C/N as Dependent			
1. log Y_H, N_W, N	0.69	0.61	0.19
2. log Y_H, log N_W, log N	0.73	0.66	0.17
3. log Y_H, N_W/N	0.53	0.37	0.04
4. log Y_H, N_W/N, log N	0.73	0.66	—
5. EA, EA^2, N_W, N	0.31	0.37	0.31
6. EA, EA^2, log N_W, log N	0.32	0.40	0.28
7. EA, EA^2, N_W/N	0.23	0.25	0.19
8. EA, EA^2, N_W/N, log N	0.32	0.40	0.28
2. C/N as Dependent			
9. Y_H, N_W, N	0.59	0.15	0.11
10. Y_H, N_W/N	0.44	—	—
11. YH, N_W/N, N	0.59	0.14	—
12. EA, EA^2, N_W, N	0.31	0.26	0.21
13. EA, EA^2, N_W/N	0.21	—	—
14. EA, EA^2, N_W/N, N	0.31	0.26	0.21

varies from one stratum to another. Also, this table points up that the truncation procedure may be absorbing too large a proportion of the total number of observations (at least for these samples), since it leaves for the regression analysis in each city less than half of the total observations and, in the case of Lima, less than one-fourth.

Turning to the regression analysis, an overall summary of the goodness of fit (R^2 adjusted for degrees of freedom) obtained from fitting the functions discussed in the preceding section to the truncated set, A_{22}, is provided in Table 2 for each of the three cities.[10] Blanks in the table indicate that a particular function was not fitted to the data in that city, usually because prior results had suggested it to be very unlikely that the particular function would be better than other functions already fitted.

On the whole, the goodness of fit obtained with these functions tends to be quite satisfactory, especially since these are cross-section data with individual households as the unit of observation. The logarithmic functions for the two Colombian cities yield much higher goodness of fit when income of the head of the household is included as a variable, but the reverse is true for Lima. The reason in the latter case would seem to be a peculiarity of the data for individual family members for Lima, a peculiarity that has appeared in other analyses of these data, and which suggests that in the case of Lima income of the family is a much more reliable indicator than income of individual earners.

To examine the importance of individual variables, the values of the coefficients of the variables included in four of these functions are presented in

[10]Estimation was by ordinary least squares. Generalized least squares was not used because the truncation procedure eliminated most of the heteroscedasticity which usually characterizes a variable such as consumption *per capita*. Because of the truncation procedure, goodness-of-fit measures have to be interpreted very cautiously, since they refer to a residual subset of the total observations.

TABLE 3

RESULTS OF SELECTED REGRESSIONS ON *PER CAPITA* CONSUMPTION, SAMPLE A

	Bogota				Medellin				Lima			
	log C/N Dependent		C/N Dependent		log C/N Dependent		C/N Dependent		log C/N Dependent		C/N Dependent	
Variable	Fn. 2	Fn. 8	Fn. 9	Fn. 12	Fn. 2	Fn. 8	Fn. 9	Fn. 12	Fn. 2	Fn. 8	Fn. 9	Fn. 12
Constant	5.87‡	12.30‡	132.2‡	136.5‡	8.51‡	12.04‡	162.5‡	75.73*	13.74‡	13.49‡	722.6‡	610.0‡
log Y_H	0.767†				0.597‡				0.026*		−0.0000	
log N	−0.756†	−0.549†			−0.904‡	−0.656‡			−0.523‡	−0.480‡		
log N_W	−0.008				−0.020				−0.038			
N_W/N		−0.0045*				0.0038				0.0007		
Y_H			0.151‡				0.0024					
N			−15.05‡	−10.97‡			−10.1‡	−11.84‡			−41.3‡	−39.8‡
N_W			0.588	−9.93			−6.26	2.89			−14.1	−3.09
E_1A		0.036†		3.37		0.017		1.38				
E_2A		−0.0006†		−0.053*		0.0000		0.0040		−0.0000		−0.0054
E_3A		0.0041		0.696		0.019		1.64		−0.017		−2.08
E_4A		0.0000		−0.0004		−0.0000		0.0012		0.0002		0.029
E_5A		0.014		1.40		0.042‡		3.44		−0.0042		0.279
E_1A^2		−0.0000		0.003		−0.0004*		−0.033				
E_2A^2		0.027*		1.77		0.052‡		4.73†		0.0000		−0.0012
E_3A^2		−0.0002		0.014		−0.0004		−0.039		0.0001		0.033
E_4A^2		0.041‡		5.03‡		0.050†		5.15*		0.020†		12.0†
E_5A^2		−0.0004*		−0.047*		−0.0003		−0.043		−0.0002*		−0.132*
R^2adj.	0.73	0.32	0.59	0.31	0.66	0.40	0.15	0.26	0.17	0.28	0.11	0.21

*Significant at 0.10 level; †Significant at 0.05 level; ‡Significant at 0.01 level.

Table 3 for each of the three cities. The functions selected are the best of those containing income and the best of those not containing income among the logarithmic functions and the corresponding arithmetic functions, in other words, Functions 2, 8, 9, and 12.

As suggested by the previous table, the income variable is highly significant in the logarithmic function for Bogota and Medellin but not for Lima. Also, in the case of Medellin, income in the linear arithmetic form is not statistically significant at even the 0.10 level.

Table 3 also brings out the fact that family size dominates the number of wage earners in all instances, so much so that the latter variable is not statistically significant if family size is included. The dependency ratio also does not seem to have much influence, being significant at the 0.10 level in only one case, Function 8 for Bogota.

The education-age interaction variables show mixed results. Only a few of the variables are significant at the 0.10 level or more; they are more likely to be significant in the logarithmic form; and there is clear support for the age-squared interaction with education. Indeed, in this and other specifications, the age-squared interaction with education tends to be more important than the age-education interaction variables alone. At the same time, these interaction effects are highly concentrated, suggesting that equally good results could be obtained more parsimoniously. This is supported by some empirical tests made with the data for Lima. Thus, fitting Function 8 using only the two age-education interaction terms for E_4 yields an adjusted R^2 of 0.24 compared to 0.28 for the full set. Fitting the same function using the four age-interaction terms involving E_4 and E_5 yields an adjusted R^2 of 0.27. For the linear arithmetic form, using only the age-interaction terms involving E_4 and E_5 yields an adjusted R^2 of 0.20 compared to 0.21 for the full set.

Overall, the signs and magnitudes of the coefficients seem to "make sense", whenever they are significantly different from zero. Thus, *per capita* consumption is positively associated with income of the head of the household, negatively associated with family size, number of wage earners and dependency burden, and generally tends to rise and then fall for a particular education as age rises. The few exceptions relate to the latter instance, notably to Lima, where the logarithmic function implies that for those households where the head has a high school education, *per capita* consumption declines at an increasing rate with age; in this instance the results with the linear arithmetic form make more sense. In the case of Bogota, the logarithmic results for E_3 (complete primary) and E_5 (complete secondary) are unsatisfactory, since they show consumption per head declining with age. The simple three-way distinction—no education, some primary or some secondary—would probably give as good or better results, although there are adequate numbers of families in the E_3 and E_5 classes. The same problem arises for Medellin. For the remaining groups E_1 (no education), E_2 (incomplete primary) and E_4 (incomplete secondary), income varies little with age at low schooling levels, but rises increasingly with age as the head is more educated. The three profiles are sharply separated even at the age of entering the labor force (15–20 years).

Now, how well do these functions perform in the key task of discriminating between poverty and nonpoverty households? The answer, for the same four functions covered in Table 3, is given in Table 4. This table shows for each function in each city the proportion of the households in Sample A that were correctly classified as in poverty or nonpoverty, and the proportions of households that were incorrectly classified as being in poverty when they were not and as not being in poverty when they were.

TABLE 4

ACCURACY OF CLASSIFICATION OF HOUSEHOLDS BY POVERTY STATUS, SAMPLE A

Dependent Variable	Fn no.	Independent Variables	Classification: Percent			
			Correct	Incorrectly Poor	Incorrectly Nonpoor	Base
		Bogota				
$\ln C/N$	2	$\ln Y_H, \ln N_W, \ln N$	83.5	11.2	5.3	193
	8	$EA, EA^2, N_W/N, \ln N$	79.3	11.6	9.1	193
C/N	9	Y_H, N_W, N	81.6	4.1	14.3	192
	12	EA, EA^2, N_W, N	71.5	7.2	21.4	193
		Medellin				
$\ln C/N$	2	$\ln Y_H, \ln N_W, \ln N$	85.3	3.8	10.9	161
	8	$EA, EA^2, N_W/N, \ln N$	77.2	5.7	17.1	161
C/N	9	Y_H, N_W, N	69.8	1.7	28.5	160
	12	EA, EA^2, N_W, N	74.7	3.1	22.2	160
		Lima				
$\ln C/N$	2	$\ln Y_H, \ln N_W, \ln N$	70.0	2.0	28.0	154
	8	$EA, EA^2, N_W/N, \ln N$	71.2	5.3	23.5	154
C/N	9	Y_H, N_W, N	72.2	0.9	27.0	154
	12	EA, EA^2, N_W, N	74.2	1.5	24.3	154

Considering the wide range of values of R^2 shown in Table 2 for these functions, it is rather surprising to find that the proportion of households correctly classified varies in a relatively narrow range, between approximately 70 and 85 percent. For a particular city, the best-fitting functions do tend to have the highest proportion of correct classification though differences are small. Thus, in the case of Lima Function 12 classifies slightly more households correctly than Function 9 although the former has a much higher goodness of fit.

Differences are also evident between cities. For example, more households are correctly classified by Function 2 for Medellin than by the same function for Bogota even though the latter has a higher goodness of fit. On the other hand, Function 8 is more accurate for Bogota than for Medellin although its goodness of fit is much higher in the latter case.

As a result, while the very best functions in terms of goodness of fit, those containing an income variable for the Colombian cities, do provide higher accuracy of classification, the margin of superiority is less than might otherwise have been expected. Thus, while 83 percent of the households in Bogota are

correctly classified by the logarithmic function containing income, substituting education-age interaction variables for income reduces the accuracy only to 79 percent. For Lima, the situation is actually the reverse—the functions containing education-age interaction variables do better than the functions with an income variable.

Where households are misclassified, what is the nature of the error? Table 4 indicates that as a rule by far the most frequent type of error is to classify poverty households as not being in poverty, in Lima and in Medellin. In other words, poverty households are being too frequently overlooked. In Bogota, however, the nature of the error varies with the type of the function, the tendency being for the logarithmic functions to classify too many nonpoor households as being poor and for the arithmetic functions to miss too many poor households.

Overview of Sample A

We are now ready to consider how the model works in its entirety. This is done in Table 5, which shows the accuracy of classification of the different components of Sample A and of the total for each city. Now the observations are adjusted for the different sampling rates, thereby indicating how accurate such a procedure might be if applied to the actual populations. For Statum A_{22} the best regression functions are used in each case, namely, Function 2 for the Colombian cities and Function 12 for Lima.

TABLE 5

OVERALL CLASSIFICATION OF HOUSEHOLDS BY POVERTY STATUS, SAMPLE A

Stratum	Classification: Percent			Weight in Population
	Correct	Incorrectly Poor	Incorrectly Nonpoor	
		Bogota		
A_1 (nonpoor)	95.9	4.1	—	0.068
A_{21} (nonpoor)	86.8	13.2	—	0.178
A_{23} (poor)	79.7	—	20.3	0.205
A_{22}	83.5	11.2	5.3	0.549
Total	84.1	8.8	7.1	1.000
		Medellin		
A_1 (nonpoor)	97.7	2.3	—	0.067
A_{21} (nonpoor)	85.5	14.5	—	0.177
A_{23} (poor)	75.0	—	25.0	0.253
A_{22}	85.3	3.8	10.9	0.503
Total	83.6	4.6	11.8	1.000
		Lima		
A_1 (nonpoor)	97.0	3.0	—	0.025
A_{21} (nonpoor)	87.3	12.7	—	0.300
A_{23} (poor)	66.2	—	33.8	0.352
A_{22}	74.2	1.5	24.3	0.323
Total	75.9	4.4	19.7	1.000

As is evident from this table, overall accuracy of classification is approximately 84 percent for the two Colombian cities and 76 percent for Lima. The two kinds of error are about equally frequent for Bogota, but for Medellin and for Lima by far the more frequent type of error is to classify poverty households as being nonpoor.

In all three instances the "weak" point in the model is the criteria used for stratum A_{23}, where the accuracy of classification is appreciably lower than for the other three strata. Although this is less true for Bogota than for the other two cities, it does suggest future work with this model might explore more stringent criteria for stratum A_{23}, possibly shifting more of the burden of classification to the regression models, especially since stratum A_{23} constitutes a substantial part of the total population in each of the three cities.

In any event, the results in Table 5 would seem to be far superior to what might be expected by chance allocation. For example, if on a purely random basis 40 percent of the sample households were classified as being in poverty and 60 percent as not (on the basis of this being the true distribution in the population), the expected proportion correctly classified would be 52 percent. If one sought to maximize the expected accuracy by classifying every household in the sample as being nonpoor (a ridiculous procedure from a policy point of view), the accuracy would still be only 60 percent.

6. Validation Test

From an analytical point of view, a much more meaningful test of the adequacy of the model is its application to another set of data from the same population. If the search process involved in developing and estimating a model served primarily to pick up quirks in that particular set of data, the results when the model is applied to a different set of data should be appreciably poorer than before. On the other hand, if the model is valid, the classification accuracy obtained by applying the Sample A functions to the data for Sample B should be within sampling error range of that shown in Table 5.

The test was carried out by truncating Sample B in the identical manner described for Sample A. The households in the residual stratum, B_{22}, were then classified as being in or out of poverty on the basis of results obtained by substituting the characteristics of each household in turn into the appropriate "best" Sample A function for that city, namely, Function 2 for the Colombian cities and Function 12 for Lima. The result of this process is presented in Table 6, which is an overall classification summary for Sample B exactly analogous to the classification of Sample A in Table 5.

Comparison of these two tables indicates that the model does almost as well for the validation sample as it does for the original sample. Thus, the overall classification accuracy for Bogota is 79.4 percent for Sample B compared to 84.1 percent for Sample A (the difference being not quite statistically significant at the 0.10 level); for Medellin the difference between the two classification percentages is only 0.8 percent and in Lima the classification accuracy is actually higher for the validation sample than for the original sample.

TABLE 6

OVERALL CLASSIFICATION OF HOUSEHOLDS BY POVERTY STATUS, SAMPLE B

Stratum	Classification: Percent			Weight in Population
	Correct	Incorrectly Poor	Incorrectly Nonpoor	
		Bogota		
B_1 (nonpoor)	99.0	1.0	—	0.070
B_{21} (nonpoor)	84.1	15.9	—	0.124
B_{23} (poor)	71.5	—	28.5	0.229
B_{22}	79.2	12.0	8.8	0.577
Total	79.4	9.0	11.6	1.000
		Medellin		
B_1 (nonpoor)	98.7	1.3	—	0.072
B_{21} (nonpoor)	86.4	13.6	—	0.106
B_{23} (poor)	75.7	—	24.3	0.279
B_{22}	83.6	6.1	10.3	0.543
Total	82.8	4.8	12.4	1.000
		Lima		
B_1 (nonpoor)	99.8	0.2	—	0.025
B_{21} (nonpoor)	82.8	17.2	—	0.282
B_{23} (poor)	75.1	—	24.9	0.341
B_{22}	71.8	4.3	23.9	0.352
Total	76.7	6.4	16.9	1.000

Further examination of these tables reveals an interesting pattern, which is not unexpected in view of the two distinct analytical steps involved in the application of this model. The first of these steps, the truncation process, involves the imposition of certain criteria but without applying any parameters derived from one sample to the other sample. In such a case, there is no reason why we should expect the results from one sample to be uniformly different than the results from the other sample, assuming of course that both samples are from the same population. However, the second procedure, the regressions applied to Stratum A_{22}, does involve such restraints, in the sense that data from Sample B are classified on the basis of parameters *estimated* from Sample A. In this case, one could hardly expect the results from Sample B to be any better than were obtained for the sample (A) from which the parameters were originally estimated. Indeed, to the extent that search bias is present, it should show up when we compare the classification accuracies for Stratum A_{22}.

Comparison of Tables 5 and 6 corroborates this interpretation. For the three initial strata, where households were classified by the truncation procedures, either sample is equally likely to be superior. In fact, of the nine such strata in the three cities covered, the accuracy of classification is higher by more than one percentage point three times for Sample B, three times for Sample A, and is virtually identical the other three times.

By contrast, for Stratum A_{22} where the regression procedure is applied, lower accuracy is obtained from Sample B in all three cities. It is a pleasant surprise to note that in all three instances the differences are small, namely, 4.3 percent for Bogota, 1.7 percent for Medellin and 2.4 percent for Lima; even the biggest of these differences is no larger than one standard error of the difference between the relevant percentages. The inference would therefore seem to be that these models have been influenced minimally by search bias.

7. Concluding Comments

This paper summarizes the results of an exploratory study utilizing a model combining truncation with regression analysis for pinpointing poverty households. The highly parsimonious nature of this model (using only family size and number of wage earners in addition to either income or an education-age combination) suggests that even more effective results could be obtained through further work on this methodology.

The fact that this model yields almost identical results when the data are separated into analysis and validation samples supports the validity of this approach. Also, for both samples, the results are far superior to what would be expected on the basis of chance allocation. At the same time, the pattern of errors is by no means random, the most frequent type of error being to classify poverty households as being nonpoor.

Especially significant is the fact that the accuracy of the discrimination is nearly as high when financial variables are excluded as when they are included. From a survey point of view this means that there is a good deal of flexibility in deciding what variables to collect in a study seeking to pinpoint poverty households. Unlike other types of studies, income and financial information do not seem to possess the importance that might otherwise be ascribed to them. Since virtually equivalent results are possible without such information, unless these variables are desired for other purposes, consideration can be given to excluding them altogether, thereby avoiding the antagonism that such questions frequently generate. (Indeed, any loss of efficiency of the study due to not seeking such information may be more than compensated by the better response that may thereby be obtained, both in terms of the quality of the information and response rates.)

The results of this study also suggest that superior results than were obtained here should be possible if more complete information is obtained on the employment status of the different members of a household and on the contributions of each to household income. This does not mean necessarily seeking information on exact amounts but rather obtaining information on the type of activity of each member if employed and on sources of income.

Still another type of information that would be very useful in such studies, and to which relatively little attention has been given, is subjective information on the current status and satisfaction of the key household members. Thus, an evaluation of the "normalcy" of income and other data provided at the time of the interview could be very helpful in adjusting for transitory elements in the financial data. Also of use both from this point of view and from the point of

view of policy, would be information on the subjective evaluation of the household members of their satisfaction with their current status and their expectations for the future, both for their own status and for that of their children. Admittedly, data of this type are not always easy to obtain, and considerable controversy surrounds their validity, but such questions can only be answered if attempts are made to obtain this information, perhaps by several different means.

Overall, the data collection process has to interact at this stage with model development in an iterative manner. The present results suggest a number of variables which should be tested within the framework of this model and on which data need to be collected. These will include both variables which can be directly affected by policy and those which cannot, though in fact even a nonpolicy variable such as age can be useful in the development of programs for at least mitigating the effects of poverty. Once these data are collected, the variables can be tested by means of the model, which in turn may lead to ideas for additional types of data to improve the model still further.

In closing it should be stressed that only a single model has been tested for discriminating between poverty and nonpoverty households. Although this model seems on an *a priori* basis to be a reasonable one, and receives strong empirical support, it is only one of numerous models that might be tested.