INCOME-GENERATING FUNCTIONS IN A LOW INCOME COUNTRY: COLOMBIA

BY GARY S. FIELDS

Cornell University

AND

T. PAUL SCHULTZ

Yale University

Income generating functions are statistical tools used to explain income inequality and other economic outcomes and behavior. These functions are often associated with a strict human capital framework, but they need not be. Instead, they may be viewed as a reduced form equation summarizing the relationship between income and various personal and locational characteristics. Following this latter interpretation, we develop the regression and analysis of variance approaches to income generating functions and estimate them empirically using micro-economic data from one low income country, Colombia. Proceeding to increasingly parsimonious specifications of income generating functions, insights are gained into the structure of incomes in Colombia.

1. INTRODUCTION

Income generating functions are statistical tools used to explain differences in personal incomes, which may be interpreted as a framework accounting for income inequality, and may be employed to infer the effect of income opportunities on a variety of economic and demographic behavior. These functions relate personal (or family) incomes to characteristics which are thought to have a predetermined effect on the level of labor earnings. Public access to individual responses from large representative household surveys and samples of population censuses provides economists with a flexible data base for more accurately fitting the parameters to these income functions, subject to the usual caveats of the quality of the data and problems of bias due to response selectivity (Heckman, 1976; Olsen, 1981). These income functions assist in the more adequate evaluation of the partial association between personal income and other factors underlying the distribution of income, such as location by geographic region or factor market, ownership of land and physical capital, and distinctions among workers by industry, occupation, sex and ethnic group.

In this paper, we report income-generating functions using two procedures: regression and analysis of variance (ANOVA). These two procedures are complementary in that decompositions of total inequality by ANOVA can also be represented by parallel regression functions. We adopt the variance of the logarithms of personal money income as a measure of aggregate income dispersion. Standard analysis of variance procedures (Fisher, 1938; Scheffe, 1959; Kim and Kohout, 1975) are then applied to a large Colombian sample to decompose the log variance into main effects of education, age and region, interaction effects, and residual within-cell variances. Equivalent regression techniques become the

basis for then testing the sequence of restrictions implicit in widely-used but highly-simplified earnings functions proposed by Mincer (1974). We also explore the usefulness of stratifying by occupation (employer/employee) and by type of residence (rural/urban) in the Colombian context. A brief recapitulation of results concludes the paper.

2. Questions, Methodology and Data

This paper presents the results of estimating a linear model of income determination in Colombia. Two closely-related linear models are used: analysis of variance (ANOVA) and multiple regression.

Analysis of variance decomposes overall income variance (or the variance in the logarithm of income) into within-category and between-category components, measures the direct contribution of each set of categories to total variance, and tests the marginal statistical significance of these effects.[1] In comparison with other decomposable measures of inequality, specifically the Theil index of inequality and the Gini coefficient, ANOVA has three advantages: (i) generally-accepted tests of statistical significance are available for ANOVA and not for the other decomposition procedures;[2] (ii) the log variance measure of inequality attaches greater importance to the relative income status of the poor;[3] and (iii) because of ANOVA's equivalence to multiple regression, effects of various influences on income may be quantified.

The data for this paper are taken from the 14th Colombian Census of Population (October 1973). A four percent Public Use Sample of Census returns was provided to us by the National Statistical Office (known by its Spanish acronym, DANE, 1977). The number of usable cases was 777,000.

To determine income, the Census asked: "What was your income in pesos last month?" Thus, one cannot distinguish labor earnings from other forms of non-labor income. We distinguished several types of income recipients. One category is day workers (jornaleros), wage laborers (obreros), and salaried employees (empleados), whom we call "employees." Self-employed (trabajadores independientes) and employers (patrones) are combined in a second category called "employers." Other types of workers (principally domestic servants and unpaid family workers) comprise a residual category which is omitted for various reasons.[4]

[1]ANOVA procedures have long been used to analyze experimental or quasi-experimental data, but on the problem of determining income and income inequality, work is more recent; see Schultz (1965), Langoni (1972, 1975), Fishlow (1973), and Chiswick (1976).

[2]This advantage is less important in our work than in most other income distribution research because of our exceptionally large sample.

[3]Champernowne (1974).

[4]Unpaid family workers are excluded for lack of income data. Domestic servants and other unspecified workers were also omitted from this analysis in the belief that income in kind, both food and lodging, makes up a substantial but unmeasured fraction of their labor earnings. Also omitted from the working sample are individuals who reported themselves employed but having zero incomes (about one percent), presumably because they failed to respond to the Census income question. Finally, women are excluded because they are thought more likely to work irregularly and part-time, which complicates any interpretation of age as a proxy for labor market experience; one-sixth of the Colombian labor force sample are women. Also, correction for selectivity bias would be unavoidable if we included women in our analysis.

For the group of "employees," the income reported includes for the most part labor earnings. For "employers," though, the income reported in the Census is likely to include not only returns to their labors and their entrepreneurial talents but also returns on their land and reproducible wealth. For this reason, we initially treat the two groups separately, and later analyze the pooled sample.[5]

A working sample of 16,695 male employees and 6,090 employers is selected, as every fifth such individual in the 4 percent DANE sample; our analyses deal with income, educational level, age, residence by rural/urban and Department,[6] and type of employment. Extensive cross tabulations of these data including also women are found in Fields and Schultz (1977), and are available from us upon request. In what follows, we present the results of regression and ANOVA for male samples. The work reported here extends an earlier study of interregional inequality in Colombia (Fields and Schultz, 1980).

3. Empirical Evidence

Our dependent variable is the natural logarithm of monthly income in pesos; the unemployed reporting no income are attributed one peso per month. The explanatory categories are education, age, and place of residence. Four educational categories are distinguished: none, primary (some or all), secondary (some or all), and higher (some or all). There are seven age categories: 10–19, 20–24, 25–29, 30–34, 35–44, 45–54, 55 and over. Three place-of-residence variables are analyzed. One is rural/urban. The second is Department of Residence (23 in number), and a third is a grouping of the Departments into six relatively cohesive regions. In most instances, results are reported here for brevity only for male employees, though the employer sample produced similar results. Later in this section, the two samples will be pooled as one test of their similarity.

Analysis of Variance: Interactive Model

A main effects model with two-way interactions is reported in Table 1. This extends Fields and Schultz (1980), which considered only non-interactive specifications. The first column displays the simple association between the logarithm of income and each set of explanatory categories; these numbers are comparable to the simple zero order correlation in the two category case. All of the main effects are by conventional statistical standards highly significant at significance levels surpassing 0.001.[7]

[5]In interpreting the results (see C. Chiswick, 1975), it should be recognized that large numbers of Colombian workers shift from employee to employer status over the life cycle. In our sample, 14 percent of the income recipients in the 20–24 age group are employers, whereas the fraction rises to 47 percent at age 55–64. Consequently, if employers earn more (less) than employees, the within-employment = type age-income profiles would systematically understate (overstate) the actual increase in income anticipated by a representative worker.

[6]Colombia is divided into 22 departments, and the special district of Bogota. A number of frontier territories and small islands (less than 2 percent of the population) are excluded from the Census sample.

[7]Given the very large sample size, virtually any basis for grouping the data according to personal, demographic, economic, social or geographic information would reduce the standard error of estimate sufficiently to satisfy the F test for statistical significance. This test starts to have discriminating power when many degrees of freedom are consumed to parameterize interaction effects.

TABLE 1

ANALYSIS OF VARIANCE WITH INTERACTION EFFECTS: MALE EMPLOYEES

	Zero Order Correlation (1)	Proportion of Variance Explained (2)	*F* Ratio Marginal (3)	df (4)
Main Effects				
Education (4)*	0.48	0.122	1,103	3
Age (7)*	0.31	0.064	286	6
Region (6)*	0.21	0.011	58	5
Rural/Urban (1)*	0.37	0.027	738	1
Covariance		0.126	—	—
Main Effects, Total		0.350	631	15
Two Way Interactions				
Education × Age		0.005	7.76	18
Education × Region		0.003	5.45	15
Education × Rural/Urban		0.005	45.7	3
Age × Region		0.003	2.35	30
Age × Rural/Urban		0.011	48.0	6
Region × Rural/Urban		0.009	48.8	5
Covariance		0.007	—	—
Two Way Interactions, Total		0.043	15.0	77
Main Effects and Interaction Effects, Total		0.393	115	92
Logarithm of Income				
Mean		6.52		
Variance		1.52		
Sample Size		16,542		

*Number of explanatory categories in parentheses.
Note: All effects statistically significant at 0.001 level.

There are two ways of interpreting the relative importance of these effects. Column (2) reports the proportion of the variance in the logarithms of income directly explained by each set of explanatory categories. Column (3) reports the marginal *F* ratio, which deflates the explained variance by the number of categories considered and expresses the resulting reduction in standard error of estimate as a ratio to that anticipated from a random set of categories in a normally distributed population.

For employees, education provides the most information in predicting personal incomes, in the sense of explaining directly 12 percent of the log variance. Its statistical significance is also greatest with an *F* equal to 1,103. The one-way rural/urban distinction accounts for 2.7 percent of the log variance and is attributed an *F* of 735. The seven age categories account for 6.4 percent of the log variance in incomes and receive an *F* of 286. The regional distinctions, though still highly significant by conventional standards, explain less than might have been anticipated given the prominence accorded interregional variation in studies of income distribution in Colombia. The six regions account directly for 1.1 percent of the log variance with an *F* ratio of 58. A little more than one-third

of the variance of the logarithm of income is explained by these four sets of main effects. The explanatory power of this model in Colombia is high compared with similarly parsimonious models for the U.S. (Mincer, 1974) and for other low income countries (Fishlow, 1972; Langoni, 1975).

Exploring covariation among the explanatory variables in other ANOVAs not reported here, we find that the direct effect of age is not greatly influenced by the inclusion of various regional distinctions, varying narrowly from 6.4 to 7.2 percent of the explained variance. When the rural-urban distinction is considered, the direct effect of education is 12.9 percent, but education's main effect rises to 19.4 percent when the six regions are included but rural/urban is omitted. Simultaneously, the covariance effect falls by more than half, confirming the strong association between education, age and the rural-urban categories.

One interesting pattern emerges in the interactions. Of the interactions that appear to be relatively important (i.e. F's exceed 40), all involve the rural-urban distinction. This confirms one's intuitive sense that rural and urban labor markets in Colombia differ in more respects than in income level (i.e. in the main effect or intercept)—they may differ also in structure and problems of measurement, such as those caused by the omission of income-in-kind or relative price variation. Further work on the rural/urban distinction is reported below.

The 15 main effects explain 35 percent of the variance of the logarithms of income among Colombian workers. The 77 two-way interactions add only an additional 4 percentage points of explanatory power. These interaction effects meet conventional statistical standards of significance, yet, relatively little predictive accuracy, about one-tenth, is gained by the inclusion of five times the number of two-way interactions as there were original main effects. For this reason, interaction effects are deemed of secondary importance in Colombia, and are not considered further here.

Quantification of Personal and Regional Effects

In order to evaluate the *magnitude* of various categorical effects (as distinct from their mere existence, which is established by ANOVA), regression estimates are helpful. Table 2 presents regression results estimated for the same sample of male employees as was used in the ANOVA in Table 1. The regression and ANOVA models are comparable, but they are not exactly equivalent for two reasons: the ANOVA in Table 1 includes two-way interactions whereas the regression model in Table 2 does not, and the regression in Table 2 uses 23 departments as the geographic breakdown rather than 6 regions.[8] All effects are expressed proportionately from geometric means, since the income generating function is in semi-logarithmic form.

In regression (1) the coefficients on the 22 department of residence dummy variables are included, but for brevity only their joint statistical significance is reported; they together account for ten percent of the variation in incomes. Adding the rural/urban dummy variable in regression (2) suggests that the department income differences may be mostly a reflection of rural/urban

[8]The ANOVA in Table 1 used the cruder geographic information to keep the number of two-way interactions within computationally feasible limits.

TABLE 2

INCOME FUNCTIONS BASED ON CATEGORICAL DATA: UNRESTRICTED AND RESTRICTED SPECIFICATIONS FOR MALE EMPLOYEES

(t ratios reported in parentheses beneath coefficients)

Explanatory Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Education:							
(Deviation from Primary)							
None			−0.300	−0.303	−0.453		
			(14)	(13)	(20)		
Secondary			0.709	0.735	0.926		
			(33)	(34)	(43)		
Higher			1.73	1.76	1.96		
			(42)	(42)	(46)		
Years						0.140	0.173
						(54)	(63)
Age:							
(Deviation from 25–29)							
10–19			−0.575	−0.582	−0.624		
			(21)	(21)	(21)		
20–24			−0.242	−0.243	−0.260		
			(8.94)	(8.79)	(9.16)		
30–34			0.144	0.140	0.155		
			(4.82)	(4.57)	(4.94)		
35–44			0.237	0.231	0.257		
			(8.83)	(8.41)	(9.12)		
45–54			0.218	0.220	0.252		
			(7.14)	(7.04)	(7.88)		
55+			−0.024	−0.0399	−0.0329		
			(0.67)	(1.09)	(0.87)		
Years						0.0996	
						(28)	
Years^2						−0.00112	
						(25)	
Experience							
Years							0.0743
							(34)
Years^2							−0.00108
							(28)
Zone:							
Rural-Urban		−0.799	−0.438	−0.550		−0.535	−0.535
(Rural = 1)		(40)	(23)	(30)		(30)	(30)
Departments: (22)							
Joint F-tests*	90.0	36.7	34.0				
(d.f.)	(22, 16, 520)	(22, 26, 519)	(22, 16, 510)				
Intercept	6.28	6.71	6.54	6.63	6.41	4.29	5.12
R^2	0.1071	0.1852	0.3531	0.3236	0.2865	0.3204	0.3529
Standard Error of Estimate	1.167	1.115	0.994	1.015	1.043	1.018	1.014
Sample Size	16,542	16,542	16,542	16,542	16,542	16,542	16,542

*Rather than report 22 coefficients on all 22 department dummy variables, the overall joint F is reported for their inclusion in the regression and the appropriate degrees of freedom underlying the F test.

compositional differences. Inclusion of the age and education characteristics of the worker in regression (3) accounts also for much of the rural/urban differences.

The age and education effects follow a standard pattern. Workers age 10–19 earn 58 percent less than workers age 25–29. Incomes rise with age in the cross section, peaking between 35 and 44, at which age incomes are on average 24 percent higher than for those in the late twenties. Employees with no schooling earn 30 percent less than those with some primary education, while those with secondary education earn 71 percent more, and those with higher education earn 1.7 times as much as those with a primary education. Overall the education and age categories alone account for about 29 percent of the log variance of incomes in regression (5).

Covariation among regional and individual characteristics was observed in the ANOVA of Table 1. This leads us to expect that some part of the differences in income across regions would be associated with differences in the educational attainment of the labor force and with age structure. In particular, since disproportionately many well-educated persons have migrated to urban areas (Schultz, 1971; Fields, 1979), the unadjusted rural/urban differential likely overstates the average differential for persons of given education. The empirical question is by how much. The rural/urban differential declines from 0.80 (regression 2) to 0.44 with the inclusion of age and education (regression 3), indicating almost half of the income differences between rural and urban male employees can be explained by these rough indicators of skill and experience. The average absolute magnitude of the departmental deviations, however, do not decrease when adjusted for age and education; they increase slightly from 0.21 to 0.23.

Comparing regressions with and without department of residence, 32.3 percent of the log variance of incomes is explained by 11 categorical age, education, and rural/urban variables (regression 3), whereas the additional 22 department variables in regression (4) increase the proportion explained only to 35.5 percent. Conversely, these 22 regional variables decrease the standard error of estimate by only 2 percent. Thus, recognition of department of residence, while informative, complicates the simple linear model without adding substantially to its predictive precision. Although a standard F ratio test would suggest the need to include regional effects, and indeed a multitude of interaction effects (Table 1), the search for a simpler income generating function appears to justify neglecting geographic detail even in a country such as Colombia where interregional disparities are emphasized.[9] However, dropping the rural/urban distinction would *not* be justified, judging by the regression coefficients in (5) compared with (4).

Earnings Functions and Simplifying Restrictions

Research on income and its determinants commonly expresses education and age in years rather than as dummy categorical variables and then fits various

[9]The marginal F ratio test of any restriction on the main effects model is not likely to be accepted given the large size of the working sample (16,680) relative to the number of parameters being fitted (32 in regression 5). See Griliches (1976).

functional forms.[10] Two restrictions are considered here that transform the age and schooling categories from the unrestricted estimation of nine parameters (six age and three education dummy variables) to three (age, age squared and schooling). To maintain comparability with the ANOVA calculations, schooling and age are measured by the mean years in each category.[11] Moving from the unrestricted main effects model without department effects (regression (4) in Table 2) to the restricted model in regression (6) the R^2 decreases 0.3 percentage points and the standard error of estimate increases by 0.003.[12] An alternative specification assumes a quadratic in post-school experience rather than age (Mincer, 1974). When direct information on experience is unavailable, a proxy is often used equal to age minus years of schooling completed minus age of school entry (in Colombia, seven). The earnings function specified in terms of a quadratic in this proxy for experience is estimated in regression (7). This transformation of age not only fits the income data better than the quadratic in age (regression 6), but it even accounts for the Colombian data better than the unrestricted model in age (regression 4). Beyond its better fit, a further advantage of the experience transformation is that the estimated coefficient on the schooling variable can be interpreted in the human capital framework as a rate of return to education. The experience transformation of age provides a theoretical justification for the specification of the earnings function, without impeding its fit to the Colombian data.

It can be shown from regression (5) in Table 2 that the parameterization of education in years is roughly consistent with the unrestricted parameter estimates, which imply a relative gain in income per year of schooling from primary, secondary, and higher education of 14, 19 and 16 percent, respectively. When relative gains per year to education range within such narrow bounds over the spectrum of educational levels in a society, and when the experience quadratic fits income data as well as it does in Columbia, there appears to be little explanatory power lost by adopting the simple specification of the income generating function derived by Mincer (1974).[13]

Comparing Employees and Employers

We began by dividing by employment-type (employees vs. employers) in order to reduce probable bias that would arise by mixing returns to wealth of

[10]Other efforts to search statistically for the best functional forms for the dependent and independent variables in the earnings function have been based on various data sets for the U.S. See Heckman and Polachek (1974) and Welland (1976).

[11]The mean years of schooling completed by employees with "primary education" is 3.3; the "secondary education" category of employees has 8.2 years; and the "higher education" category of employees report 14.9 years. With respect to age the midpoints of the categories are treated as the means from age 20 to 54, and the average age of the youngest and oldest age category are set equal to 17 and 62 years respectively.

[12]Even in this case the F ratio test rejects the restriction given the sample size.

[13]The regressions in Table 2 are based on categorical information (e.g. knowledge that a particular individual is in age category 35–44) rather than more exact, virtually continuous data (e.g. the individual is 43 years old). This was done in order to compare parallel ANOVA and regression specifications. To determine how much information was lost by the use of categorical data, continuous age and education information was also considered (reported below). Based on the continuous variables, the proportion of variance explained tends to increase about three percentage points.

the self-employed with returns from labor. As Fishlow (1972, 1973) has argued in his study of the distribution of income in Brazil, it seems likely that education in particular would be strongly associated with the control of capital, ownership of land, and access to influential institutions and people. Consequently, education's association with income could capture not only an effect of skills on labor's productivity, but also the influence of family social status and wealth on personal income.[14] These may differ as between employees and employers.

Separate earnings functions for male employees and employers in Colombia are presented in Table 3, using alternately age and experience. The two sets of results are similar in regression coefficients and proportions of variance explained. Given these findings, we combined the employee and employer samples and estimated income-generating functions for the pooled sample with separate intercepts. The regression results are shown in columns (5) and (6) of Table 3, the ANOVA results in Table 4. In regression (6) the coefficient on the employer dummy variable is 0.07, indicating that employers received about 7 percent higher incomes than employees,[15] holding constant for the direct effects of age,

TABLE 3

INCOME FUNCTIONS REGRESSIONS BASED ON CATEGORICAL DATA: MALE EMPLOYEES AND EMPLOYERS

(t ratios reported in parentheses beneath coefficients)

Explanatory Variable	Employees		Employers		Employees and Employers	
	(1)	(2)	(3)	(4)	(5)	(6)
Years of Schooling	0.140	0.173	0.160	0.179	0.145	0.174
	(54)	(63)	(30)	(32)	(61)	(69)
Age	0.100		0.105		0.0984	
	(30)		(13)		(32)	
$Age^2/100$	−0.112		−0.115		−0.110	
	(25)		(12)		(28)	
Experience		0.743		0.0756		0.0746
		(34)		(14)		(36)
$Experience^2/100$		−0.108		−0.107		−0.108
		(30)		(13)		(31)
Rural/Urban	−0.535	−0.535	−0.830	−0.825	−0.613	−0.612
(Rural = 1)	(30)	(30)	(21)	(21)	(36)	(36)
Employer/Employee					0.0828	0.0701
(Employer = 1)					(4.76)	(4.03)
Intercept	4.285	5.123	4.22	5.22	4.295	5.136
	(74)	(166)	(27)	(62)	(77)	(173)
R^2	0.320	0.326	0.276	0.278	0.304	0.309
Standard Error of Estimate	1.018	1.014	1.348	1.346	1.119	1.243
Sample Size	16,542	16,542	6,090	6,090	22,632	22,632

[14]For examination of international aspects of education in Columbia, see Kugler (1975), Fields (1976), and Berry and Urrutia (1976).

[15] According to Chiswick's (1975) formulation of the earnings function for the self-employed, the regression coefficient on the self-employment variable can be interpreted in the human capital framework as $-\log_e \alpha$, where α is the labor share of income received by the self-employed. Among Colombian male self-employed and employers, these estimates suggest approximately 93 percent of their incomes are imputed returns to their labor, holding constant for age, education and region effects.

TABLE 4

ANALYSIS OF VARIANCE WITH MAIN EFFECTS: POOLED MALE EMPLOYEES AND MALE EMPLOYERS

Main Effects	Zero Order Correlation	Proportion of Variance Explained	*F* Ratio Marginal
Education (4)*	0.47	0.113	1,293
Age (7)*	0.29	0.054	308
Region (6)*	0.29	0.017	57
Rural/Urban (1)*	0.37	0.020	669
Employer/Employee (1)*	0.08	0.001	46
Covariance		0.135	—
Total Explained		0.340	554
Logarithm of Income:			
Mean	6.57		
Variance	1.80		
Sample Size	22,632		

*Number of explanatory categories in parentheses

Note: All effects are statistically significant at 0.001 level.

education, department, and rural/urban, the effects of which are quite similar for employers and employees.[16] In the ANOVA, the employer/employee variable accounts directly for only 0.1 percent of the log variance in incomes among Colombian men.[17]

These results from employee/employer comparisons indicate that the two groups do not have a different *structure* to their earnings functions; rather, the *level* (intercept) of the function is seven percent higher for employers. This contrasts with much larger differences found in Brazil using a slightly different range of employment categories (Fishlow, 1973; Langoni, 1975). The relative effects of education and experience in Colombia are somewhat more pronounced among employees; as an explanation of incomes among employers, region and particularly rural/urban location are more important. Nonetheless, pooling the two employment groups does not alter the form of the earnings function greatly, other than in the intercept.

Comparison of Urban and Rural Areas

Colombia's population is about evenly divided between urban and rural locations. In all statistical tests reported above, urban/rural location appeared as a significant determinant of incomes. Furthermore, when interactions were

[16]The standard error of estimate is increased by only 0.5 percent when the restriction is imposed that all of the regional dummy variables, schooling, age, and the age quadratic be identical for both employers and employees. This set of 13 parameter restrictions on the general ANOVA model implies an *F* ratio of 10.7 with 12 and 22,808 degrees of freedom. These restrictions would not be accepted by standard statistical conventions, yet in terms of predictive adequacy of the model the pooled results are nearly as good as the stratified results.

[17]The effect is statistically significant by the conventional *F* test, but with a sample of 22,000 plus, this is hardly surprising.

allowed for, substantial covariance appeared between urban/rural and other explanatory variables. This suggests that the explanatory contribution of the other independent variables (education, age, and department) may differ as between rural and urban areas. We now explore those differences.

The most straightforward way of testing for rural/urban differences is to divide the population into two groups, rural and urban, and to examine the structure of income determinants in each. It is also desirable to distinguish between employees and employers. Analysis of variance results are presented in Table 5 for these four strata of the male population. To quantify the differences between education and experience effects for employees in rural and urban areas, Table 6 reports two fully interactive income regressions; the first specification neglects differences in personal incomes by department, and the second specification allows for such differences. Education and experience variables are measured in Table 6 continuously rather than categorically, increasing slightly the explanatory power of these simple income generating functions. Several results are noted:

(1) The relative explanatory power of education, age, and department differs greatly between the rural and urban samples.

(2) In urban areas, for both employees and employers, *education* and *age* are the principal explanatory variables; department plays a minor role. More specifically, for urban employees, of the 30.9 percent of the log variance explained, 17.6 percent is directly explained by education, 9.6 percent by age, and 1.2 percent by department. Likewise, for urban employers, the respective figures are 25.4 percent (total), 17.2 percent (education), 4.6 percent (age), and 1.2 percent (department).

(3) In rural areas, for both employees and employers, *department* is the principal explanatory variable; education and age play minor roles. More specifically, for rural employees, of the 17.7 percent of the log variance explained, 13.8 percent is directly explained by department, 1.7 percent by age and 1.6 percent by education. Likewise, for rural employers, the respective figures are 27.3 percent (total), 23.7 percent (department), 0.7 percent (age), and 1.5 percent (education).

(4) Given that education and age are important determinants of income in urban but not in rural areas and that interdepartmental differences are important in rural but not in urban areas, we might expect interregional movements in labor to respond to these differential rewards. Education raises income proportionately more in urban areas, 19 percent per year of schooling compared with 8 percent in rural areas, and the overall level of income is also higher in urban than rural areas at all levels of education. Accordingly, educated persons have the strongest incentive to leave rural areas and migrate to the cities. Less-educated individuals also have an incentive to migrate from low-income departments, and insofar as the high-income departments generally include major cities, their migration may also be rural-to-urban. Research in Colombia has already established that net migration flows in the 1951–1964 intercensal period were closely associated with differences between municipal daily agricultural wages and the relatively common structure of urban earnings (Schultz, 1971), and that gross lifetime migration patterns among departments recorded in the

TABLE 5

ANALYSIS OF VARIANCE WITH MAIN EFFECTS: MALES, STRATIFIED BY URBAN/RURAL AND EMPLOYEE/EMPLOYER

	(1)		(2)		(3)		(4)	
	Urban Employees		Rural Employees		Urban Employers		Rural Employers	
Main Effects	Proportion of Variance Explained	*F* Ratio, Marginal	Proportion of Variance Explained	*F* Ratio, Marginal	Proportion of Variance Explained	*F* Ratio Marginal	Proportion of Variance Explained	*F* Ratio Marginal
Education (4)*	0.170	846	0.016	40	0.172	443	0.015	27
Age (7)*	0.096	245	0.017	20	0.046	59	0.007	6
Department (23)*	0.012	8.0	0.138	45	0.012	4.2	0.237	57
Covariance	0.032	—	0.005	—	0.025	—	0.014	—
Total Explained	0.309	153	0.177	41	0.254	64	0.273	47
Logarithm of Income								
Mean	6.88		5.90		7.20		5.93	
Variance	1.45		1.03		1.97		2.45	
Sample Size	10,591		5,951		3,928		2,162	

*Number of explanatory categories in parentheses

Note: All effects statistically significant at 0.001 level

TABLE 6

INCOME FUNCTIONS BASED ON CONTINUOUS DATA: MALE EMPLOYEES WITH RURAL INTERACTIONS

(*t* ratios reported in parentheses beneath coefficients, and degrees of freedom beneath *F* test statistics)

Explanatory Variables and Joint *F* Tests	(1)	(2)
Years of education	0.192	0.187
	(65)	(65)
Years of experience	0.0920	0.0916
	(35)	(36)
Years of experience squared	−0.135	−0.134
(÷100)	(27)	(28)
Department (22) effects*	—	8.75
Joint *F* test (d.f.)		(22, 16, 490)
Overall intercept	4.82	4.74
Rural Interactions with the Following:		
Years of education	−0.112	−0.111
	(14)	(14)
Years of experience	−0.0560	−0.0575
	(12)	(12)
Years of experience squared	0.0818	0.0843
(÷100)	(10)	(11)
Department (22) effects*	—	14.42
Joint *F* test (d.f.)		(22, 16, 490)
Intercept	0.460	0.613
	(7.36)	(6.44)
Rural interactions on only education, experience, and intercept	304	—
Joint *F* test (d.f.)	(4,16534)	
All rural interactions including department	—	25
Joint *F* test (d.f.)		(44, 16, 490)
R^2	0.3385	0.3799
Standard Error of Estimate	1.004	0.974
Sample Size	16,542	16,542

*Rather than report 22 coefficients on all 22 department dummy variables, the overall joint *F* is reported for their inclusion in the regression and the appropriate degrees of freedom underlying the *F* test.

1973 census remain strongly associated with personal income levels (Fields, 1979).

(5) Comparing employees and employers in rural areas, the income structures are different. Although the two groups have similar means (5.90 and 5.93, respectively), the logarithmic variance of income is much greater for employers (2.45) than for employees (1.03). This larger variance is accounted for, at least in part, by greater interdepartmental variation among employers,[18] (particularly the self-employed—not shown). This suggests that the labor market for landless rural workers (farm laborers and non-agricultural employees) is relatively

[18]Compare the relative explanatory power of department for the two groups.

uniform geographically, but the distributions of wealth and returns on that wealth in farming and ranching are not. Presumably, these differences are associated with the size distribution of landholdings, altitude and climate conditions, and specific cropping and tenure patterns, but these speculations remain to be explored in detail.[19]

(6) Rural and urban labor markets in Colombia differ both in level of income and in income structure, i.e., returns to education and experience. The differential rates of technical change in the two sectors, disparate rates of capital formation and modernization, effective protection, and rapid rural-to-urban migration have undoubtedly contributed to these distinct income structures in rural and urban areas. The precise ways in which these and other forces operate over time to determine incomes are a challenge to future research. Several salient predetermined factors affecting personal incomes, including education, experience and possibly region, can readily be held constant by conventional statistical procedures to help disentangle how many remaining factors determine the personal distribution of income. The proposed simplified income generating functions estimated in this paper do not appear to impose arbitrary restrictions on the personal income data from the 1973 Colombian Census. Parallel analyses of micro data sets from other countries which include good information on personal incomes should advance our understanding of how demographic, education and institutional factors alter income inequality.

4. Summary and Conclusions

A four-percent sample of the 1973 Colombian Census of Population is analyzed to clarify the determinants of income and income variance. Among male employees, education, age, region, and rural/urban differences in income are distinguished using decompositions of the log variance of income (ANOVA) and by parallel regression techniques.

The ANOVA results support the hypothesis that education, age, region, and rural/urban location contribute significantly in accounting for the log variance of income in Colombia. By standard statistical conventions, the four-way classification by educational attainment is much the more important, while the single urban/rural dichotomy is next in importance per degree of freedom used. The seven age categories are generally more significant statistically than the six, or twenty-three, regional categories.

The fifteen parameters used to model the main effects of education (3 parameters), age (6) region (5), and urban/rural (1) account for one-third of the log variance in incomes of employees (and somewhat less of those of employers). Interaction effects represented by 77 additional parameters were found to account for only an additional 3 to 4 percent of the log variance of incomes.

[19] A review of the literature on rural income distribution in Colombia turned up many tabulations but no suitably disaggregated data on the correlates of rural wage structure. The literature reports that average income increases with the size of the landholding, some regions are richer and experience more rapid growth than others, and returns to education are lower in rural areas than in urban areas. The interested reader is referred to the book by Berry (forthcoming) and the studies by Berry and Soligo (1980).

That is, a proportionate model of income determination which is linear in the variables and ignores interaction effects does almost as well as a more complex specification with interactions among all of these variables.

The goodness of fit of the earnings function was then examined, with the restriction that (1) the effect of years of schooling on income is proportionate at all levels of education, and (2) life cycle proportionate variation in income can be approximated by a quadratic in age or years of post-schooling experience. As compared with the general model, the restricted earnings function results in only a small (0.1 percent) increase in the standard error of estimate when based on the same categorical age information. The standard error is actually reduced when the experience transformation of age and schooling is used in the regression. Replacing the categorical age and schooling data by the underlying virtually continuous information available from the census increases the explanatory power of this simple human capital specification further.[20]

The employer and employee samples were then pooled. The employment-type distinction was found to contribute only one-tenth of one percent to the explanation of the log variance in incomes, even though employers received 7 percent more income than employees, other things equal. This is because the income variation within employee and employer groups is so much greater than the variation between them. This contrasts with similar calculations performed on Brazilian census data (Fishlow, 1973; Langoni, 1975) in which employment position was a major explanatory variable that also reduced the magnitude of schooling's effect on the logarithm of income. Estimating a single income generating function for employees and employers in Colombia would not appear to do violence to the patterns of income distribution in that country.

Finally, urban and rural samples were analyzed separately. The simple linear model does somewhat better in explaining income variance in urban than in rural areas. But more importantly, pronounced differences in the structure of incomes in the two areas were encountered: urban incomes vary largely with education and age, while rural incomes vary with region. The urban labor market is relatively similar across the 23 departments, suggesting an integration and homogeneity in returns to schooling and experience that would hardly be expected, given the rugged terrain separating the many growing urban centers of Colombia. Conversely, the large regional variation in incomes in the rural sector implies additional important factors affecting income have been omitted and perhaps also that these labor markets are in disequilibrium.

For rural employers the interregional income differences are undoubtedly due in part to agricultural factor endowments other than labor, such as the quality and quantity of land owned, and the size distribution of these holdings. But among employees (landless rural workers) these inter-regional differences in labor income within education/age groups might be explained by differences in the relative price levels across regions, particularly in basic foodstuffs, and the availability of nonmonetized household incomes which are neglected in the census definition of monthly money income (Lecaros, 1979). Probably more important is underlying disequilibrium among rural labor markets scattered

[20]See footnote 13.

through numerous relatively isolated areas of Colombia. Persisting differences in the level and educational structure of incomes between rural and urban areas are also notable in Colombia.

Initially, the proposition was advanced that income-generating functions are a useful tool for describing and understanding income variation in low income countries. The expanding availability of sample information on household economic and demographic characteristics can support more specific and detailed inquiries into the causes of income variation and how various forms of household behavior adapt to the evolving distribution of income opportunities that occur with development. The Colombian census sample prepared by DANE was opened to the public less than three years after the census was conducted. Studies, such as this, of the income data from the census sample have already yielded descriptive and prescriptive information for Colombians. Related investigations of labor supply behavior, migration, fertility, child mortality, marriage behavior and the distributional effects of effective protection have already relied on this valuable public census sample. The Colombian example should allay the fears of skeptics about the capacity of the statistical offices of low income countries to produce prompt and reliable household samples from their population and housing censuses. The example of the Colombian Statistics Department (DANE) should be widely followed, both in industrialized and developing countries, and perhaps more effort expended in the future to coordinate population and agricultural censuses in order to illuminate some of the unresolved puzzles noted here in interpreting the distribution of income among persons in rural areas.

Bibliography

Berry, R. Albert, *The Development of Colombian Agriculture*, forthcoming.

Berry, R. Albert and Miguel Urrutia, *Income Distribution in Colombia*, New Haven, Yale University Press, 1976.

Berry, R. Albert and Ronald Soligo, *Economic Policy and Income Distribution in Colombia*, Boulder Colorado, Westview Press, 1980.

Champernowne, D. W., A Comparison of Measures of Income Distribution, *The Economic Journal*, December, 1974.

Chiswick, Carmel U., Determinants of Earnings from Self-Employment, Development Research Center, The World Bank (mimeo.), September, 1975.

———, Income Distribution in Thailand: Application of the Theil Index to Income Inequality, Development Research Center, The World Bank, Series B-2, 1976.

DANE (Departmento Administrativo Nacional de Estadistica), *La Poblacion en Columbia 1973, Muestra de Avance*, Bogotá, Colombia, 1977.

Fields, Gary S., Education and Economic Mobility in Colombia, Economic Growth Center, Yale University, Center Discussion Paper No. 237, revised 1976.

———, Lifetime Migration in Colombia: Tests of the Expected Income Hypothesis, *Population and Development Review*, June, 1979.

Fields, Gary S. and T. Paul Schultz, Sources of Income Variation in Colombia: Personal and Regional Effects, Economic Growth Center, Yale University, Center Discussion Paper No. 262, June, 1977.

———, Regional Inequality and Other Sources of Income Variation in Colombia, *Economic Development and Cultural Change*, 28, 3, April, 1980.

Fisher, R. A., *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd, 1938.

Fishlow, Albert, Brazilian Size Distribution of Income, *American Economic Review*, May, 1972.

———, "Brazilian Income Size Distribution—Another Look," *Dados*, 1973, II.

Griliches, Zvi. Wages of Very Young Men, *Journal of Political Economy*, 2, August, 1976.

Heckman, James and Solomon Polachek, Empirical Evidence on the Functional Form of the Earnings-Schooling Relationship, *Journal of the American Statistical Association*, June, 1974.
Kim, Jae-On and Frank J. Kohout, Analysis of Variance and Covariance, Chapter 22 in Norman H. Nie et al., ed., *Statistical Package for the Social Sciences*, New York, McGraw Hill, 1975.
Kugler, Bernardo, Influencia de la Educación en los Ingresos del Trabajo: El Caso Colombiano, *Revista de Planeacion y Desarrollo*, 1975.
Langoni, C., Distribuicao da Renda e Desenvolvimento Economico do Brasil, *Estudos Economicos*, October, 1972.
———, Income Distribution and Economic Development: Brazilian Case, Paper presented at the Econometric Society World Congress, Toronto, August, 1975.
Lecaros, C. G. de, Regional Development: Education and Income in Rural Colombia, unpublished Ph.D. dissertation, Stanford University, February, 1979.
Mincer, Jacob, *Schooling, Experience, and Earnings* New York, Columbia University Press, 1974.
Scheffe, H., *The Analysis of Variance*, New York, Wiley, 1959.
Schultz, T. Paul. The Distribution of Personal Income: Case Study of the Netherlands, unpublished Ph.D. dissertation, MIT, Cambridge, Mass., 1965.
———, Rural-Urban Migration in Colombia, *Review of Economics and Statistics*, May, 1971.
Welland, J. D., Cognitive Abilities, Schooling and Earnings: The Question of Functional Form, McMaster University, Working Paper No. 76-14, November, 1976.