PREDICTING INTERNAL MIGRATION: AN APPLICATION OF REGRESSION TECHNIQUES *

M. E. El Attar and R. R. Hocking
Mississippi State University

INTRODUCTION

The state of Mississippi has been exposed to relatively high net out-migration, especially before 1960. For the state as a whole, net migration estimates amounted to -433,584 (-19.9 percent) and -267,394 (-12.3 percent) for the two decades of 1950 to 1960 and 1960 to 1970, respectively (U. S. Bureau of the Census, 1962 and 1971a). Such substantial net out-migration is likely to promote marked changes in the overall composition of the state's population; and this justifies the investigation of migration correlates.

OBJECTIVES OF THE STUDY

The present study builds on earlier works (El Attar, 1970, 1973, 1974; El Attar and Tarver, 1972; El Attar and Windham, 1973; El Attar and Steinman, 1975) and is one of several directed toward determining the factors to which internal migration is most related. Specifically, this paper examines the feasibility of predicting county net migration (dependent variable) by using a small number of predictors or independent variables.

SELECTION OF PREDICTORS

Selection of predictors in this study was achieved in two stages: (1) initial selection, and (2) final selection.

Initial Selection of Predictors

The initial selection of predictors in this study is based on a theoretical model developed by the senior author (El Attar, 1970; El Attar and Tarver, 1972). According to this model, the process of migration involves two factors: (1) an epiphenomenal behavior produced by socio-cultural compulsions which influences individuals according to the society's division of labor and (2) the incentives to move, which vary in time and space. The model uses as a basic framework the concepts of attraction, repulsion and compulsion. Within the context of this framework a number of socioeconomic and demographic variables were selected as predictors, namely changes in civilian employment, unemployment, school enrollment, population under five and 65 years and over, and median income of families and unrelated individuals.

Description of Variables and Data

Data on which this study is based come principally from Mississippi decennial censuses and other publications of the U. S. Bureau of the Census. Specifically, these sources provide comparable data on the dependent and independent variables. Table 1 provides a numerical summary on each of the selected variables in the decennial censuses of 1950, 1960 and 1970. A brief description of the trend in each variable is given below.

Net migration estimates.--Table 1 demonstrates that the state was subjected to a substantial net out-migration during the two decades from 1950 to 1970, as specified in the introduction. However, it should be indicated that the rate of net out-migration between 1960 and 1970 declined by slightly over one-third of its level in the preceding decade (166,190 persons).

Of the 82 counties in Mississippi only six experienced net in-migration (37,624 or 11.2 percent) during the 1950's.[1] In the 1960's although the number of counties that gained population through migration increased to nine,[2] the number of in-migrants was numerically smaller (36,003), but proportionately greater (17.6 percent), than for the previous decade.

Employment.--An examination of Table 1 reveals that, while total employment of persons 14 years old and over in Mississippi decreased by 4.8 percent (34,512) between 1950 and 1960, between the latter year and 1970, total employment increased by 6.2 percent (42,360). It also should be noted that there were significant changes in the number of counties gaining employment as well as the employment change. During the two decades from 1950 to 1970, the number of these counties increased from 21 to 48 and the numerical (not percent change) in their employment increased from 43,039 (18.2 percent) to 74,436 (17.8 percent), respectively.

Unemployment.--Unemployment data in this study include all persons 14 years old and over who are "not 'at work' but looking for work" (U. S. Bureau of the Census, 1963). Since the age coverage of unemployed persons in the 1970 census county data is limited to persons 16 years old and over, unemployed persons 14 and 15 years old have been estimated and added in order to make the 1970 data comparable to those of 1960.[3]

Table 1 shows that, while total unemployment of persons 14 years old and over in 1960 (39,284) increased by over 51 percent of its level in 1950 (25,976), between 1960 and 1970, unemployment decreased by 1.4 percent (547), and the number of counties experiencing reduction in their unemployment level increased from seven to 42 counties.

School enrollment.--Data on school enrollment are for persons 5 to 34 years old in 1960 and 1970.[4] Table 1 shows that, while school enrollment for the state increased by 19.5 percent (100,035) between 1950 and 1960, between the latter year and 1970, school enrollment increased by only 5.6 percent (34,064). A different pattern is representative of the changes in county school enrollment. An examination of the counties with positive changes in school enrollment during the two decades between 1950 and 1970 reveals that school enrollment in these counties decreased[5] as a percentage of total enrollment from 77.4 (397,290) to 59.6 (385,793) percent; their number dropped from 56 to 34 counties.

Median income.--Changes in county real median incomes of families and unrelated individuals provide the fourth predictor in this paper. The median income data as used in this research are comparable. Moreover, in order to enhance the comparability, county median income data for 1959 and 1969 were measured at constant purchasing power, taking 1967=100 (U. S. Bureau of the Census, 1972b).[6]

Table 1 shows that both money and real median incomes of families and unrelated individuals in the state increased significantly since 1949. However, it is quite clear that the increase for the second decade is not as high as that for the first one, especially when the real value is considered. The same can be said of the changes in real median incomes experienced by counties. All counties showed increases in their real median incomes for families and unrelated individuals; with the change being greatest in De Soto County, 193.6 percent (from $2,109 in 1959 to $6,193 in 1969), and lowest in Pike County, 22.4 percent (from $3,064 in 1959 to $3,751 in 1969).

Population under five and 65-plus years old.--Table 1 indicates that the proportion of population below age five and 65 years old and over to the total population yield two opposing trends in which the magnitude of the former declined whereas that of the latter increased in each of the three censuses since 1950. Youngesters' proportion declined from 13.0 to 9.5 percent in 1950 and 1970, respectively; proportion of the olds increased from 7.0 to 10.0 percent in same dates. The percent changes between the decades convey a different message, which constitute a declining trend for both the young and old populations. Between 1960 and 1970, positive changes were the characteristic for only six counties[7] in the case of population under 5 years of age, whereas negative changes characterized the population 65 years old and over in eight counties.[8]

DATA ANALYSIS AND FINAL SET OF PREDICTORS

A linear model for predicting migration is considered. Because of the variation in the socioeconomic and demographic conditions of the counties, all predictors were expressed relative to their 1960 level. Thus, for example, the employment variable, one of the six predictors, denotes the 1960 to 1970 change in a county employment divided by the 1960 county employment. The dependent variable, migration, is the 1970 county net migration rate.

For this transformed data, the linear model

$$m_i = a + \sum_{j=1}^{6} b_j x_{ij} + e_i \qquad i = 1...82$$

was fitted using standard regression techniques. Analysis of the residual plots and lack of fit error suggest that the model is reasonable. The results of the regression analysis are shown in Table 2. The independent variables are ordered as they entered the model in a forward selection model (see Draper and Smith, 1966). The cumulative R^2 reveals the increase in R^2 achieved by adding one variable at a time assuming all preceding variables are kept in the model. Thus, for example, with the first four variables in the model, adjoining unemployment and median income increases R^2 by only 0.0013. The regression coefficients and standard errors shown are those for the equation with all six independent variables included. Also shown are the simple correlations between migration and each of the independent variables and the analogous partial correlations. It is of interest to note that the signs on these two correlation coefficients are opposite for each of the last two variables; however none of these are significant at the 5 percent level, hence these differences may be attributed to sampling variability.

In an effort to provide further information on the roles of the independent variables as predictors of migration, an analysis was made on all $2^6 = 64$ possible subset regressions. It was observed that the optimum subsets of each size agreed with those recommended by the stepwise procedure (and also those recommended by backward elimination). This provides evidence of the lack of any serious problem with colinearity among the independent variables.

The possibility of deleting some variables without substantially affecting the ability of the resulting equation to predict migration is suggested by the cumulative R^2 values. To provide further support for variable deletion and to suggest how many variables might be deleted we considered the C_p-statistic proposed by Mallows (1964, 1966). If we let p denote the number of terms in the subset equation (number of variables plus one for the intercept), then

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

Here RSS_p is the residual sum of squares for the p-term equation, $\hat{\sigma}^2$ is the residual mean square for the equation with all variables included and n is

the number of observations.

The nice feature of the C_p-statistic for subset selection is that it provides simple guidelines for determining the number of variables to be included. In general, small values of C_p are desirable with the added restriction that C_p should not be substantially larger than p to avoid bias. In particular, attention is directed at the minimum C_p if $C_p < p$.

In Table 3 we show the values of C_p and also the residual mean square, RMS_p, for the best subset of each size for p = 2,. . ., 7. Note that the minimum C_p occurs for p = 5 and suggests the deletion of variables 5 and 6, namely unemployment and median income. Inspection of other subsets reveals that none of the next best subsets have values of C_p less than 6. Thus the subset consisting of variables 1 to 4 seems quite reasonable. (For a detailed discussion of subset selection criteria, see Hocking: 1972 and 1976).

Table 3 provides the estimates of the coefficients and their standard errors for the subset equation ignoring the possible bias introduced by deleting variables 5 and 6.

SUMMARY AND CONCLUSION

This study examined the relationship between net migration and changes in a selected set of socioeconomic and demographic variables in Mississippi. Also, it investigated the feasibility of predicting county net migration by using a linear regression model and the C_p-statistic. Application of the C_p-statistic provided useful criteria for evaluating the best subset of predictors. According to these criteria, income and unemployment could be deleted from the predicting model, hence, reducing the number of predictors to four: changes in (1) employment, (2) population under five years of age, (3) population 65-years old and over, and (4) school enrollment. This economy in the number of predictors will reduce the cost and effort of migration prediction.

These findings are of interest to the students of social sciences and statistics as well as to policy planners and decision makers of community development. Future migration prediction should consider age, sex, and racial migration differentials. Accounting for these differences is expected to enhance reliability of the prediction.

TABLE 1. NUMERICAL SUMMARY OF VARIABLES

Item		Year and Amount			Change	
		1950	1960	1970	1950-60	1960-70
Net migration	No.	-433,584	-267,394		166,190	
	Rate	-19.9	-12.3		7.6	
Employment	No.	716,851	682,339	724,699	-34,512	42,360
	Rate	96.5	94.6	94.9	-4.8	6.2
Unemployment	No.	25,976	39,284	38,737	13,308	-547
	Rate	3.5	5.4	5.1	51.2	-1.4
School enrollment	No.	513,335	613,370	647,434	100,035	34,064
	Pct.	52.7	57.3	56.8	19.5	5.6
Median income: Money	Value	$1,032	$2,324	$4,658	$1,292	$2,334
	Pct.	-	-	-	125.2	100.4
Median income: Real	Value	$1,446	$2,661	$4,243	$1,215	$1,582
	Pct.	-	-	-	117.7	68.1
Population under 5 years	No.	283,349	278,412	209,606	-4,937	-68,806
	Pct.	13.0	12.8	9.5	-1.7	-24.7
Population 65 years and over	No.	152,964	190,029	222,320	37,065	32,091
	Pct.	7.0	8.7	10.0	24.2	17.0

Source: U. S. Bureau of the Census, 1952, 1961a, 1961b, 1962, 1963, 1971a, 1971b, and 1972a.

TABLE 2. REGRESSION ANALYSIS FOR COUNTY NET MIGRATION WITH SIX PREDICTORS, MISSISSIPPI: 1960 - 1970

Number of Variables	Predictors by Order of Entry and Intercept	Cumulative R^2	Regression and Intercept	Standard Errors	Correlation	
					Simple	Partial
1	Employment	.8840	.3639*	.0604	.9402*	.5708*
2	Population < 5	.9136	.3613*	.0654	.8981*	.5380*
3	Population 65+	.9199	.1236*	.0778	.7592*	.1804*
4	School enrollment	.9238	.0612*	.0354	.7352*	.1960*
5	Unemployment	.9247	.0127	.0111	-.0394	.1306
6	Income	.9251	-.0150	.0222	.0845	-.0780
	Intercept		-.0664			

*Significant at .05.

TABLE 3. SUBSET ANALYSIS

Number of Variables	P	Variables Included	C_p	RMS_p (10^2)	Reg Coeff for P = 5	STD Error
1	2	1	38.2	0.231	.348	0.058
2	3	1, 2	10.5	0.174	.356	0.034
3	4	1, 2, 3	6.2	0.164	.147	0.063
4	5	1, 2, 3, 4	4.4	0.158	.068	0.065
5	6	1, 2, 3, 4, 5	5.0	0.158	-	-
6	7	1, 2, 3, 4, 5, 6	7.0	0.159	-	-
Intercept					-.079	

FOOTNOTES

*The research on which this paper is based is part of MAFES Population Project No. 4004.

[1]Hancock, Harrison, Hinds, Jackson, Lowndes, and Rankin.

[2]De Soto, Hancock, Itawamba, Jackson, Pearl River, Prentiss, Rankin, Stone, and Tishomingo.

[3]Table 121 (U. S. Bureau of the Census, 1972a) provides "percent in labor force" by age and sex. A preliminary estimation was obtained by multiplying the percent in labor force in ages 14 and 15 by the population in these ages, for males and females separately. The resultant figures were added to civilian labor force 16 years old and over and civilian employment was subtracted from this figure, yielding the preliminary county unemployment estimates. Finally, the preliminary figure for each county was revised by having it multiplied by the ratio of total unemployment for the state, as given

in the census, (38,737), to the total aggregated from the preliminary county estimates (38,202).

[4]For 1950, school enrollment data are for persons 5 to 29 years old. For comparison purpose the authors believe that the extended age coverage for the 1960 enrollment data is not significant. This statement is supported by the fact that the magnitude of enrollment for persons 30 to 34 years old in 1960 in Mississippi represented only 0.9 percent of total enrollment. The 1970 school enrollment data for counties are made comparable to those of 1960 by applying a procedure analogous to the one used in adjusting county unemployment data, as indicated in Footnote 3 above.

[5]It must be noted that part of this decline is related to a decline in population fertility. In 1970 the total fertility rate of Mississippi was 2.9 births per female in the childbearing age (15 to 44), a reduction of about one-third over the 1960 rate of 4.3 per female (El Attar, 1975: 1).

[6]The purchasing power of the dollar was adjusted by using the three consumer price indices 1.401, 1.145, and .911 for 1949, 1959, and 1969, respectively.

[7]De Soto, Hancock, Itawamba, Jackson, Prentiss, and Rankin.

[8]Carroll, Holmes, Humphreys, Issaquena, Jefferson, Noxubee, Quitman, and Tunica.

REFERENCES

Draper, N. and H. Smith, 1966. Applied Regression Analysis. New York: John Wiley & Sons, Inc.

El Attar, M. E., 1970. "Migration and Occupational and Industrial Changes in Georgia, 1950-1960." Athens: The University of Georgia, unpublished Ph. D. dissertation.

El Attar, M. E. and J. D. Tarver, 1972. "A Theoretical Model for Predicting Internal Migration in the United States." 1972 Proceedings of the Social Statistics Section, American Statistical Association, pp. 227-232.

El Attar, M. E., 1973. "Migration and Changes in the Industrial Structure of Mississippi: A Multivariate Analysis." Paper presented at the annual meeting of the Rural Sociological Society, University of Maryland, Maryland, August 23-26.

El Attar, M. E. and Gerald O. Windham, 1973. "Migration and Occupational Change in Mississippi: A Multivariate Analysis." Proceedings: Rural Sociology Section. Association of Southern Agricultural Workers, 16 pp., processed.

El Attar, M. E., 1974. "Migration in Relation to Some Selected Socio-Economic Factors in Mississippi." Rural Sociology in the South, 1974.

_____, 1975. Mississippi Population by County, Age, Sex, and Color: 1960, 1970, and Projections to 1980, MAFES Bulletin 815, Mississippi State.

El Attar, M. E., and D. L. Steinman, 1975. "Internal Migration Correlates in Mississippi," Rural Sociology in the South, (forthcoming).

Hocking, R. R., 1972. "Criteria for Selection of a Subset Regression: Which One Should be Used?" Technometrics, 14 (November), 967-970.

_____, 1976. "The Analysis and Selection of Variables in Linear Regression." To appear in Biometrics (February).

Mallows, C. L., 1964. "Choosing Variables in a Linear Regression: A Graphical Aid." Paper presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7-9.

_____, 1966. "Choosing a Subset Regression." Paper presented at the Joint Statistical Meeting, Los Angeles, California.

U. S. Bureau of the Census, 1952. U. S. Census of Population: 1950. Vol. II, Characteristics of the Population. Mississippi. Washington, D. C.: U. S. Government Printing Office.

_____, 1961a. U. S. Census of Population: 1960. General Population Characteristics, Mississippi. Washington, D. C.: U. S. Government Printing Office.

_____, 1961b. U. S. Census of Population: 1960. General Social and Economic Characteristics, Mississippi. Washington, D. C.: U. S. Government Printing Office.

_____, 1962. Current Population Reports, Series P-23, No. 7, Components of Population Change, 1950 to 1960, for Counties, Standard Metropolitan Statistical Areas, State Economic Areas, and Economic Subregions. Washington, D. C.: U. S. Government Printing Office, Table 1, p. 35.

_____, 1963. U. S. Census of Population: 1960. Vol. I, Characteristics of the Population. Mississippi. Washington, D. C.: U. S. Government Printing Office.

_____, 1971a. Current Population Reports, Series P-25, No. 461, Components of Population Change by County: 1960 to 1970. Washington, D. C.: U. S. Government Printing Office, p. 36.

_____, 1971b. U. S. Census of Population: 1970. General Population Characteristics, Mississippi. Washington, D. C.: U. S. Government Printing Office.

_____, 1972a. U. S. Census of Population: 1970. General Social and Economic Characteristics, Mississippi. Washington, D. C.: U. S. Government Printing Office.

_____, 1972b. Statistical Abstract of the U.S., 1972. Washington, D.C.: U.S. Government Printing Office.