ALTERNATIVE MODELS OF INDIVIDUAL MOBILITY

Alden Speare, Jr., Population Studies and Training Center, Brown University

The decision to move is usually a complex one which involves several factors and is influenced by various characteristics of the mover and his place of residence. Any model which attempts to represent this decision making process must involve several variables. A simple type of model is one in which mobility, represented by a 0,1 variable, is viewed as a linear function of independent variables such as age, education, occupation, home ownership and duration of previous residence (see [3] and [4]). In an earlier paper [8], I tried to expand this model to include intervening variables such as residential satisfaction and a previously expressed desire to move which in turn depend on individual and residence characteristics.

The parameters of these models can be determined through ordinary multiple regression by defining the mobility variable as equal to 1 for movers and 0 for non-movers. For any subgroup of the population the mean value of the mobility variable lies between 0 and 1 and can be interpreted either as the proportion of movers in that subgroup or the probability that an individual in that subgroup will be a mover. The unstandardized regression coefficients which are obtained from the multiple regression can be interpreted as the relative contributions of the independent variables to mobility if the effects can be assumed to be linear and additive. This interpretation is especially clear in the case where all independent variables are discrete variables (for a discussion of the use of discrete variables in regression see Suits, [9] and Goldberger, [2], pp. 218-227). The expected probability that a person with given characteristics will move is simply the sum of the regression coefficients that correspond to the characteristics he possesses plus the constant:

$$\hat{y}_i = a + \Sigma\, b_k\, X_{ki}$$

where $X_{ki} = 1$ if he possesses the K^{th} characteristic and 0 otherwise. The ability to divide up the probability of moving into separate components which at least crudely represent the independent effects of different variables makes multiple regression an attractive technique for the study of individual mobility. However, the use of a dichotomous dependent variable presents at least two problems.

First, the assumption of homoscedasticity is violated since the mobility variable has a binomial distribution with asymtotic variance equal to PQ. Goldberger ([2], pp. 249-250) has suggested that this problem can be solved by introducing weights inversely proportional to the variance or $(PQ)^{-1}$. Unfortunately the true value of P is unknown and must be estimated from the observed value. When the numbers in some subgroups of the sample are small, the estimates are subject to considerable error. This is especially a problem when P is near 0 or 1 for the weight gets very large and small deviations of the observed P from the true P can have large effects on the weight used.

Another problem with this model is that for some combinations of values of the independent variables, the expected value, $\hat{y}_i$, may be either greater than unity or less than zero. Anyone accustomed to evaluating the validity of a model by its behavior at extremes would be inclined to reject this model since an event cannot have a negative probability of occurrence or a probability greater than unity. Others who are willing to mentally make the conversion of a negative probability to a zero probability might still raise the objection that to use $(y_i-\hat{y}_i)^2$ provides an unnecessary addition to the variance when $\hat{y}_i$ is less than 0 or greater than 1.

Several alternative models have been suggested for dealing with this problem. One suggestion is to use a logit transformation to limit the expected value of the dependent variable to the 0 to 1 interval (see figure 1). Thiel [10] has discussed this model at length for the case where the independent variables take on a limited number of discrete values. For 3 independent variables the proportion of cases with $y_i=1$ for each subgroup defined by a combination of values on the independent variables, f_{jkl}, is estimated by P_{jkl}, where

$$P_{jkl} = (1+e^{-L_{jkl}})^{-1}$$

and

$$L_{jkl} = a + b_1 X_{1j} + b_2 X_{2k} + b_3 X_{3l}$$

P_{jkl} equals .5 when $L_{jkl} = 0$, equals 0 when $L_{jkl} = -\infty$ and 1 when $L_{jkl} = +\infty$. The regression coefficients can be determined approximately by calculating:

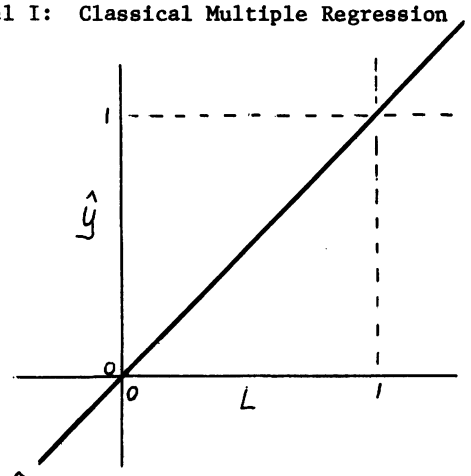
$$L_{jkl} = \ln \frac{f_{jkl}}{1-f_{jkl}}$$

and regressing this on the independent variables. Unfortunately L_{jkl} is undefined when f_{jkl} equals either 0 or 1. Thiel recommends dropping these cases from the analysis by giving them a weight equal to 0. Other procedures, such as changing f_{jkl} to what it would be if a half a case had the opposite value, are discussed by Gart and Zweifel [17]. The fact that none of these procedures provide an unbiased treatment of these cases is a weakness of this approach.

A third model is a simple modification of the classical multiple regression equation in which $\hat{y}$ is truncated at 0 and 1 (see figure 1). This can be accomplished in the computation stage by the instructions:

Figure 1
Three Alternative Multiple Regression Models

Model I: Classical Multiple Regression

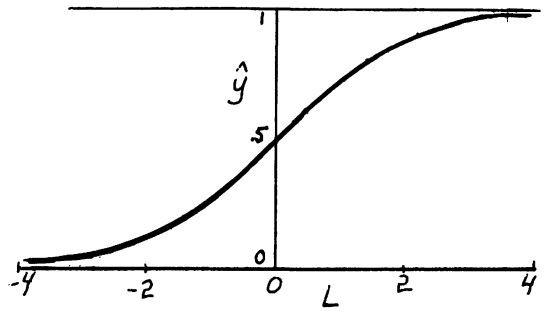


$\hat{y} = L_{jkl}$

$L_{jkl} = a + b_1 X_{1j} + b_2 X_{2k} + b_3 X_{31}$

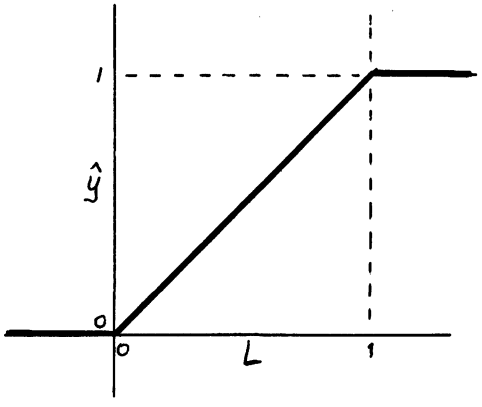
(same for all models)

Model II: Regression of Logit



$\hat{y} = (1 + e^{-L_{jkl}})^{-1}$

Model III: Truncated Classical Multiple Regression



$\hat{y} = 0$ for $L_{jkl} < 0$

$\hat{y} = L_{jkl}$ for $0 \leq L_{jkl} \leq 1$

$\hat{y} = 1$ for $L_{jkl} > 1$

if $\hat{y} < 0$, set $\hat{y} = 0$

if $\hat{y} > 1$, set $\hat{y} = 1$

Since y no longer has a continuous first derivative, the least-squares equations can not be solved directly. However, a solution can be obtained through iteration.

The choice of a model should depend on at least three criteria: (1) the ease of performing the computations involved; (2) the goodness of fit to the data, and (3) the appropriateness of the model to the theoretical assumptions about the relationships of the variables. While some complex model might provide the best goodness of fit, a simpler model might be preferable if the fit were almost as good and the model seemed more clearly in line with the theoretical assumptions.

All three of the models described above seem to be appropriate for the study of individual mobility. The classical multiple regression equation states that the probability of moving is simply a weighted sum of the values of the independent variables. This model is simple and seems reasonable in the lack of information indicating more complex relationships. The truncated multiple regression equation is similar with the exception that once the sum of factors is sufficient to predict either mobility or immobility (i.e., $\hat{y} = 1.0$ or $\hat{y} = 0$) the addition of other factors favorable to mobility (or immobility) makes no difference. The logit model is very similar to the linear model near the center. However, as one moves away from the center each additional factor has less effect on the probability of moving. The extremes of 0 and 1 can never be reached which is probably realistic in that there are always some factors which dispose a person to stay or move which are not included in any particular model.

THE DATA

The data come from interviews taken in the 1969 round of the Rhode Island Health Study and a telephone follow-up interview one year later (see [6]). The original survey included 1081 respondents who were representative of the Rhode Island population aged 21 and over and the married population of all ages. A sub-sample of 724 respondents was selected which contained all those who had ever been married, who were under 65 years of age, who were either the head of the household or spouse of the head, and who were not currently serving in the military. The age and marital status restrictions were deemed necessary because of the large variation in mobility rates with these variables (see Speare, [7]). A small number of respondents who were neither the head or spouse of the head were excluded because these persons might not be involved in the decision to move. The military were excluded because it was felt that their movement might not be entirely voluntary.

In the original interview, respondents were asked questions about their characteristics, the

characteristics of their residence, and their satisfaction with various aspects of their housing and geographical location. They were also asked whether they had any wish to move or plans to move within the next year.

Approximately one year later, these same respondents were contacted by telephone (or field interview where necessary) and were asked if they had moved. Every effort was made to obtain the follow-up interview and 95 percent of the respondents in our sub-sample were reinterviewed. In cases where the follow-up interview could not be obtained, the interviewers tried to ascertain whether or not the person had moved. Of the 724 respondents who met the criteria for this study, movement was ascertained for 711 (10 persons had died or entered institutions during the year and 3 refused to be reinterviewed.

THE ANALYSIS

A special program called NDIMA was written to perform the multiple regression allowing any of the three models to be selected. The program first tabulated the data in an N-dimensional matrix where each of N-1 independent variables and the dependent variable represents a dimension. The size of each dimension is 2, which restricted the analysis to all dichotomous variables, although variables with 3 or more categories could be handled by dividing them into two or more dichotomous variables. The least square equations were set up from the data matrix. Appropriate transformations and weights were calculated and the equations were solved through matrix inversion to yield estimates for the regression coefficients.

For models one and three, weights directly proportional to the number of cases and indirectly proportional to the estimated variance were used with the exception that when the observed proportion, f_{jkl}, was less than .05 or greater than .95 the variance for $f_{jkl} = .05$ was used. This kept the weights from getting too large and provided a finite weight for cases where $f_{jkl} = 0$ or $f_{jkl} = 1.0$. For model two, a modified form of the logit was used which was defined at $f_{jkl} = 0$ and $f_{jkl} = 1.0$ as described in Gart and Zweifel [1], p. 181:

$$L_{jkl} = \ln \frac{R_{jkl} + .5}{T_{jkl} - R_{jkl} + .5}$$

where R_{jkl} = No. of movers in cell

T_{jkl} = Total number in cell

An appropriate weight for this model is:

$$W_{jkl} = \frac{(R_{jkl} + .5)\ (T_{jkl} - R_{jkl} + .5)}{T_{jkl} + 1}$$

which is similar to the weight NPQ suggested by Thiel ([10], p. 109) when N is large, but has less bias when N is small and is defined at $f_{jkl} = 0$ and $f_{jkl} = 1$.

After a solution was obtained, new weights were calculated based on the estimated probabilities and the equations were solved again. For model three further iteration was required. A revised sum of squares was calculated by setting the estimated probability equal to 0 whenever it was negative and 1 when greater than 1. The regression coefficients were then successively incremented and decremented by a small amount and a test was made to see if the sum of squares was reduced. This process was repeated until convergence was obtained.

The following independent variables were chosen for the analysis based on previous research with the classical multiple regression model (see [8]). All independent variables were defined as of the original interview in 1969:

1. Age of the head of household: 0 = Ages 18-34; 1 = Ages 35-64.

2. Owner/Renter status: 0 = Owner; 1 = Renter or Other.

3. Duration of Residence: 0 = 0 to 4 years; 1 = 5 or more years.

4. Friends and Relatives Index. An index representing the proportion of one's friends and relatives who live in the immediate neighborhood or the same section of town. 0 = Relatively low proportion of friends and relatives; 1 = Relatively high proportion.

5. Index of Residential Satisfaction. An index made up of the weighted sum of the expressed level of satisfaction with each of eight items dealing with aspects of housing, neighborhood, and residential location. The item weights were proportional to the relative importance attributed to each item by all respondents. 0 = Relatively low satisfaction; 1 = relatively high satisfaction.

6. Wish to Move: Based on response to the question "Do you have any wish to move within the next year?" 0 = No; 1 = Yes.

THE RESULTS

The results of stepwise multiple regression for the three models are shown in Table 1. These results are based on 678 cases for which there was complete information for all of the variables. The order in which the variables are added is arbitrary although it approximates the common procedure of adding those variables which account for the greatest increase in the "explained" sum of squares first. A crude measure of the goodness of fit for each of the models is the coefficient of determination obtained by taking the ratio of explained sum of squares to the total sum of squares based on deviations of each case

TABLE 1

Results of Three Different Multiple Regression Models for Predicting Residential Mobility

	Classical Multiple	Logit Model[a]	Truncated Classical Model
A. With Three Independent Variables			
Constant	.069	-.662	.069
Wish to Move	.221	.419	.221
Age	-.056	-.226	-.056
Owner or Renter	.101	.269	.101
Coef. of Determination[b]	.168	.178	.168
B. With Four Independent Variables			
Constant	.063	-.674	.019
Wish to Move	.219	.373	.259
Satisfaction Index	.021	.121	.046
Age	-.051	-.219	-.146
Owner or Renter	.100	.246	.180
Coef. of Determination[b]	.169	.185	.187
C. With Five Independent Variables			
Constant	.073	-.650	.028
Wish to Move	.220	.390	.260
Satisfaction Index	.031	.119	.056
Age	-.046	-.158	-.096
Owner or Renter	.093	.223	.168
Duration of Residence	-.022	-.139	-.077
Coef. of Determination[b]	.171	.186	.191
D. With Six Independent Variables			
Constant	.079	-.560	.019
Wish to Move	.209	.358	.299
Satisfaction Index	.026	.096	.106
Age	-.050	-.191	-.110
Owner or Renter	.080	.227	.190
Duration of Residence	-.018	-.088	-.068
Proportion of Friends and Relatives in the Neighborhood	.020	-.123	-.110
Coef. of Determination[b]	.171	.192	.196

[a]All Coefficients have been multiplied by .25, the value of $\Delta p/\Delta \cdot L$ at p = .05

[b]Adjusted for degrees of freedom

from its expected value. This is equivalent to the correlation ratio discussed by Neter and Maynes [5]. Since any particular case must either be a mover or a non-mover, the deviations are typically large. Using this measure of goodness of fit, the logit and truncated regression models are potentially superior because it is possible to generate expected probabilities near 0 and 1 for many combinations of the independent variables.

The results are generally in agreement with these expectations. The logit model provides a better fit to the observed data for all four runs. The truncated model is the same as the classical model for three variables because none of the expected probabilities fall outside the 0 to 1 interval. However, it is superior to the classical model for four or more variables where some combinations of the independent variables require truncation.

In general, the logit model and the truncated classical model assign larger effects to the independent variables than the classical model does. This is most apparent if one compares the relative size of the last variable to be added in each run. For instance, duration of residence which is added for the five variable run has a regression coefficient of only -.022 for the classical multiple regression, but a coefficient of -.139 for the logit model and -.077 for the

truncated classical model.

The three models also differ in the decisions that are made about whether or not to add an additional variable to the model. For the classical model, none of the additions beyond three variables is statistically significant on the basis of an F-test of the increment to the explained variance. On the other hand, all of the additions to the truncated classical model, up to six independent variables, were statistically significant. Additions to the logit model were also statistically significant with the exception of duration of residence.

In summary, both the logit model and the truncated classical model are superior to the classical multiple regression model for the analysis of individual mobility. They both tend to allow for more independent variables in the model and to assign larger effects to these variables. In comparing the two models, the logit model has the advantage of providing a continuous function which can be solved exactly whereas the truncated model requires iteration. On the other hand, the results of the truncated model can be interpreted directly as components of the probability of moving attributable to different independent variables and these may be summed simply subject only to the simple transformation at the extremes. All three models encounter problems when subgroup sizes are small because of the difficulty in estimating the variance for these subgroups which is used to calculate weights. Further work is needed to establish efficient procedures for small samples.

ACKNOWLEDGMENT

This paper was prepared for the Annual Meeting of the American Statistical Association in Fort Collins, Colorado, August 23-26, 1971. The work was supported by United States Public Health Grant HS-00246 from the National Center for Health Services Research and Development. The author is grateful for the assistance of Dr. Harold Organic who directed the field study and Dr. James Sakoda who helped with the computer programming and commented on an earlier version of the paper.

REFERENCES

[1] Gart, John J., and James R. Zweifel, "On the Bias of Various Estimators of the Logit and its Variance with Application to Quantal Bioassay," Biometrika, 54, 1967, pp. 181-187.

[2] Goldberger, Arthur S., Econometric Theory, New York: John Wiley and Sons, Inc., 1964.

[3] Lansing, John B. and Eva Mueller, The Geographic Mobility of Labor, Ann Arbor: Institute for Social Research, The University of Michigan, 1967.

[4] Morrison, Peter A., "Chronic Movers and the Future Redistribution of Population: A Longitudinal Analysis," Demography, Vol. 8, 1971, pp. 171-184.

[5] Neter, John and E. Scott Maynes, "On the Appropriateness of the Correlation Coefficient with a 0,1 Dependent Variable," Journal of the American Statistical Association, Vol. 65, 1970, pp. 501-509.

[6] Organic, Harold N., and Sidney Goldstein, "The Brown University Population Research Laboratory: Its Purposes and Initial Progress," in Irving I. Kessler and Morton L. Levin (editors), The Community as an Epidemiologic Laboratory: A Casebook of Community Studies, Baltimore: The Johns Hopkins Press, 1970.

[7] Speare, Alden, Jr., "Home Ownership, Life Cycle Stage, and Residential Mobility," Demography, Vol. 7, 1970, pp. 449-458.

[8] Speare, Alden, Jr., "Residential Satisfaction as an Intervening Variable in Residential Mobility." Paper presented at the Annual Meeting of the Population Association of America, Washington, D. C., April 22-24, 1971.

[9] Suits, Daniel B., "Use of Dummy Variables in Regression Equations," Journal of the American Statistical Association, Vol. 52, 1957, pp. 548-551.

[10] Thiel, Henri, "On the Estimation of Relationships Involving Qualitative Variables," American Journal of Sociology, Vol. 76, 1970, pp. 103-154.