Housing and Society
Vol. 12, No. 3, 1985

# *OWNERSHIP AND OWNERSHIP PREFERENCE: A COMPARISON OF OLS AND LOGIT REGRESSIONS*

Sherman Hanna and Suzanne Lindamood

## *ABSTRACT*

*To assess the magnitude of the advantage of logit analysis over ordinary least squares regression, the two methods are used for similar models of ownership and ownership preference as functions of age and age squared. The extent of the differences between the two methods is shown in graphs. For most purposes, it makes little difference which method is used, so the ease of use and interpretation of OLS suggests it may often be an acceptable method.*

## *INTRODUCTION*

Housing researchers frequently use dichotomous variables and these dichotomous variables have often been used as dependent variables in multivariate models. Examples of housing-related dichotomous variables include tenure (own or rent), structure (single family house or other), satisfaction (satisfied or not) and various preference variables. Ordinary least squares (OLS) regression has been a common statistical method, but the theoretical disadvantages of using OLS have long been known (Pindyck and Rubinfeld, 1976) and alternate methods such as logit or probit have been recommended. At least eight articles in *Housing and Society* have used OLS or probit in models with dichotomous dependent variables.

Some authors imply that OLS should never be used with a dichotomous dependent variable. For instance, Kinsey and Lane (1983) state:

> Linear estimating techniques such as ordinary or weighted least squares do not yield valid test statistics with dichotomous dependent variables and there exists the possibility of estimating P(S) greater than 1 or less than 0.

Kinsey and Lane (1983) use probit for a model with a dichotomous satisfaction variable. Burgess (1982) uses probit with a model for home ownership.

---

Sherman Hanna is Professor, Human Development and Family Studies Department, Kansas State University, Manhattan, Kansas. Suzanne Lindamood is Associate Professor, Department of Clothing, Textiles and Interior Design, Kansas State University, Manhattan, Kansas.

At the other extreme, some authors continue to use OLS with dichotomous dependent variables with little discussion of the possible problems. Morris, Winter, and Sward (1984) use OLS with dichotomous dependent variables for actual single family home ownership, for single family home ownership preferences and for single family home ownership norms. They do not mention any possible statistical problems with OLS. Winter and Morris (1982) also use OLS for similar regressions, but with no mention again of possible statistical problems.

Pol (1981) uses ordinary least squares regression with ownership status as a dichotomous dependent variable, with a brief note citing Goodman(1976). Guy and Pol (1983) use OLS regressions for ownership and structure preference with the same citation of Goodman, who essentially suggests that if the dichotomous dependent variable has a mean value between 0.1 and 0.9 for all subgroups in the sample, then OLS will give a reasonable approximation to loglinear estimation.

Hanna and Lindamood (1979) use OLS regressions for ownership and structure and ownership and structure preferences. A footnote cites a study showing the similarity of OLS results to other estimating techniques. Chi (1984) discusses objections to OLS, but presents citations supporting the use of OLS because of the greater ease of interpretation. Chi uses OLS in regression models of tenure change.

Despite criticism, OLS regression has the advantages of greater simplicity, accessibility and familiarity. OLS is more commonly available in statistical programs for microcomputers. In popular statistical packages such as SPSS, OLS is much easier to use and provides a greater range of statistical output than probit or logit. Probit and logit results are more difficult to interpret than are OLS results. Probit and logit are rarely covered in first or second statistics courses, while OLS is typically mentioned in an introductory course and is covered in a second course. Use of probit or logit poses a human capital problem for many students and researchers, with the need for a substantial additional investment of time to learn probit or logit. As with any investment, one should ask if the expected benefits are worth the cost.

## *THE COMPARISON*

To test the differences between ordinary least squares and logit, comparisons are made for two models of home ownership. Graphical presentations of predicted values of the dependent variables are used to illustrate the extent of the differences between OLS and logit.

### *The Sample*

The sample used is a stratified random sample of approximately 1.8 percent of the households in the city of Montgomery, Alabama in 1976. Households were selected randomly from a computer listing of the 55,000 dwelling units in the city, with a resultant sample of 1,010. Census tracts with below-median income were sampled at about twice the rate as tracts with above-median income. Statistical analyses were weighted to make the results representative of the population. The study by Hanna and Lindamood (1979) offers additional descriptions of the sample. In order to simplify the model and the discussion, only households with married couples are included in the analyses.

*The Model*

Achieved tenure status and family norms for tenure may be related to a variety of demographic characteristics (Morris and Winter, 1978). Age of the head of the household and marital status are very important determinants of achieved tenure status. Income is probably the single most important determinant of achieved tenure status, but the impact of income is probably different at different stages of the life cycle.

Two dependent variables are analyzed: ownership status and preference for ownership. The dependent variable ownership preference is based on the answer to the question, "If you were to move in order to change your housing, would you like to own or rent?" Preference for renting is coded 0 and preference for owning is coded as 1. Ownership status is coded 0 for renting and 1 for owning. To simplify comparison of results, only one independent variable is used: the age of the head of the household. . (When additional independent variables such as income are included in the model, the similarity between OLS and logit reported in this article remains.)

Table 1 shows the percentage owning and preferring ownership for six age categories and the total subsample. The F statistics for the ANOVAs indicate that ownership status differs greatly by age, but preference for ownership differs less by age, with over 80 percent of all age groups preferring ownership. There is an inverted U pattern of preference by age.

Table 1. Mean Proportions Owning and Preferring Owning, by Age of Head (couples), Montgomery, Alabama, 1976.

| Age of Head | Owning | | Prefer Owning | |
|---|---|---|---|---|
| | % | (N) | % | (N) |
| under 25 | 25.5 | (31) | 80.1 | (31) |
| 25-34 | 57.8 | (175) | 95.1 | (174) |
| 35-44 | 75.1 | (119) | 91.5 | (119) |
| 45-54 | 80.8 | (130) | 93.4 | (129) |
| 55-64 | 84.4 | (103) | 92.3 | (101) |
| 65 and over | 82.9 | (92) | 84.3 | (88) |
| All | 71.8 | (650) | 91.4 | (641) |
| F statistic | 15.3 | (p=.0000) | 2.9 | (p=.0131) |

*Regressions with Dichotomous Variables*

Any dichotomous variable can be expressed as a probability. For instance, home ownership can be coded as 1 if a house is owned and

0 if it is rented. Although each household will be either a owner or a renter, a sample will have an ownership rate. If 60 percent of a sample owns, the probability that a household drawn at random from the sample owns is 60 percent. To discover patterns of ownership, the sample can be divided into groups, such as age, and the proportions of owners in each group can be observed.

In general, a regression model is of the form:

$$Y= a_0 + a_1X_1 + . \quad . \quad . + a_nX_n$$

With a dichotomous dependent variable, the Y actually only takes on one of two values, which typically are coded as 0 or 1, but the regression equation will "predict" a value for any observation. This value may be less than 0 or greater than 1, clearly an empirically impossible situation. Another problem is that of heteroskedasticity, which results in the estimates of the coefficients $a_0$ through $a_n$ not being efficient (Maddala, 1983).

Probit and logit are superior estimating techniques for regression models with dichotomous dependent variables, as the possibility of predicted probabilities less than 0 or greater than 1 are eliminated and the heteroskedasticity problem is reduced (Pindyck and Rubinfeld, 1976). Hrozencik (1984) provides a clear explanation of logit. Typically, there is little substantive difference between logit and probit results in terms of significance levels (Aldrich and Nelson, 1984). With the logit model, the equation estimated is:

$$Z=a_0 + a_1X_1 + . \quad . \quad . +a_nX_n$$

where:

$$Z=\log (P/(l+P))$$

in the "textbook" version of logit.

The logit procedure in the SPSSX PROBIT procedure is used in this article. The SPSSX logit uses a transformation, so that

$$Z = \log (p/(1-p) )/2 + 5.$$

The predicted probability is

$$p = Y/(1+Y)$$

where

$$Y = \exp((Z-5)*2).$$

To weight the logit analyses, the Aggregrate procedure must be used, as the Weight procedure does not work directly on logit in SPSSX. One disadvantage of logit compared to OLS is that the dependent variable is not the proportion or probability, but instead a logarithmic transformation. Therefore, interpretations of the results are more difficult than is the case with OLS. It is possible to calculate the probability P for a particular combination of values of independent variables in a logit by first calculating Z, then finding

$$P=\exp(Z)/(1+\exp(Z))$$

The estimated probability, P, is not a linear function of the independent variables. However, that does not mean that a researcher using logit should ignore possible alternate specifications of functional form.

For instance, Derrick's (1979) comparison of OLS to logit analysis specifies the work decisions of college students (dichotomized as work/not work) as a linear function of aid and potential wage. The linear specification for OLS and for logit assumes that an extra dollar increase in either independent variable has the same effect (on Z for logit) at a level of 0 as it would at the maxim value in the sample. The linear assumption is probably the cause of the large disparity between OLS and logit predictions of the probability of working, given the extreme values of the independent variables. An inappropriate assumption of linearity will cause a greater distortion in OLS than in logit, because, unlike logit, OLS is not constrained to give probabilities between 0 and 1. However, a linear specification may be inappropriate for logit. In Derrick's (1979) logit example, at a value of $2,000 for potential wage, an increase from 0 to $100 in student aid is related to almost the same decrease in the probability of working (0.0065) as an increase from $3,000 to $3,100 in student aid, so for that range of values, the logit probability is almost as linear a function of aid as is the OLS probability.

*Model Specification*

Correct model specification is crucial in statistical estimation (Aldrick and Nelson, 1984). Part of model specification involves assumptions about the relationship between each independent variable and the dependent variable, controlling for the other independent variables. One common specification is the assumption of a linear relationship. The linear specification is sometimes reasonable. If, however, there is a change in direction of the relationship, (e.g., as the independent variable increases, the dependent variable first increases, then decreases, or if the magnitude of the effect of a unit change in the independent variable depends substantially on the level of the independent variable), the linear form is inappropriate.

An alternative to linear specification is the use of dummy variables, which are based on ranges of values of an independent variable. In a regression, the number of dummy variables must always be one less than the number of categories used for the underlying independent variable. Otherwise, there would be perfect multicollinearity among the independent variables and estimation would be impossible. Age of the head of household could be included in a regression by a dummy variable coded as 1 if the head is age 65 or more and 0 if the head is under 65. If more detail were desired, five dummy variables could be used to represent 10-year intervals, with one dummy representing age 65 and over and the omitted category being under 25 (Table 1).

Examination of bivariate patterns can be helpful in deciding upon categories for dummy variables. There are disadvantages to using dummy variables. If the actual age of the head is available, information is lost on the pattern of changes, unless a large number of categories is used. If interaction effects are to be analyzed, the interactions between two sets of dummy variables (e.g., age and income) may yield a very large number of variables in the regression. If two variables are each represented by eight dummy variables, the interactions between the two variables would yield 64 additional variables. For a variable such as age, some regularity of pattern may be expected based on typical family life cycles and patterns of income. There is some loss in explanatory power in using dummy variables when one could make an alternate specification of the functional form which incorporates knowledge of the general shape of the relationship (Aigner, Goldberger and Kalton, 1975).

The pattern in Table 1 for both ownership and ownership preference shows first an increase, then a decrease with age. As discussed previously, a linear specification may be inappropriate even for logit analysis. For relationships involving age, family size or income, the quadratic functional form (linear term plus the square of the variable) may be appropriate. For instance, a logit analysis of ownership as a function of the age of head results in the relatively straight line shown in Figure 1. The logit results for the linear functional form are:

$$\underset{(t=6.4)}{Z=0.02132 * Age} + \underset{(t=31.3)}{4.54236}, \text{ Chi square}=109.6, \text{ d.f.}=64, p=0.000.$$

The logit analysis of ownership as a function of age and age squared (Table 2) clearly fits the pattern shown in Table 1 much better than the logit analysis of age only. .PP Considering ownership as a function of age and age squared, the results of OLS and logit analyses are presented in Table 2. The corresponding graph is presented in Figure 2.

The significance tests are very similar. The difference in the t statistic for the constant terms is mainly due to the difference in the dependent variables, which for OLS is actual ownership, while for the SPSSX logit procedure, the dependent variable is

$$\log (p/(l+p))/2+5$$

where p is the proportion of owners for each age by year. Prediction of ownership rates by either OLS or logit gives very similar results.

The largest difference between the OLS and the logit predictions is at age 85, with an OLS prediction of 0.63 and a logit prediction of .68, a difference of 0.05 in the ownership rate. For middle-age ranges, the differences are very small. The predicted ownership rate peaks at age 59 with OLS (0.86) and age 60 with logit (0.85). The $R^2$ for the OLS regression is slightly higher than Efron's $R^2$ for the logit.

OLS and logit analyses of ownership preference are also fairly similar in terms of significance tests (Table 3) and predictions (Figure 3). Ownership preference peaks at age 43.4 with OLS (0.94) and at age 43.7 with logit (0.94). The largest difference between the OLS and logit predictions is at age 85, with 0.72 for OLS and

**FIGURE 1 LOGIT ANALYSES OF OWNERSHIP RATE -- COMPARISON OF MODEL WITH LINEAR TERM FOR AGE OF HEAD TO MODEL WITH LINEAR AND QUADRATIC TERMS FOR AGE**

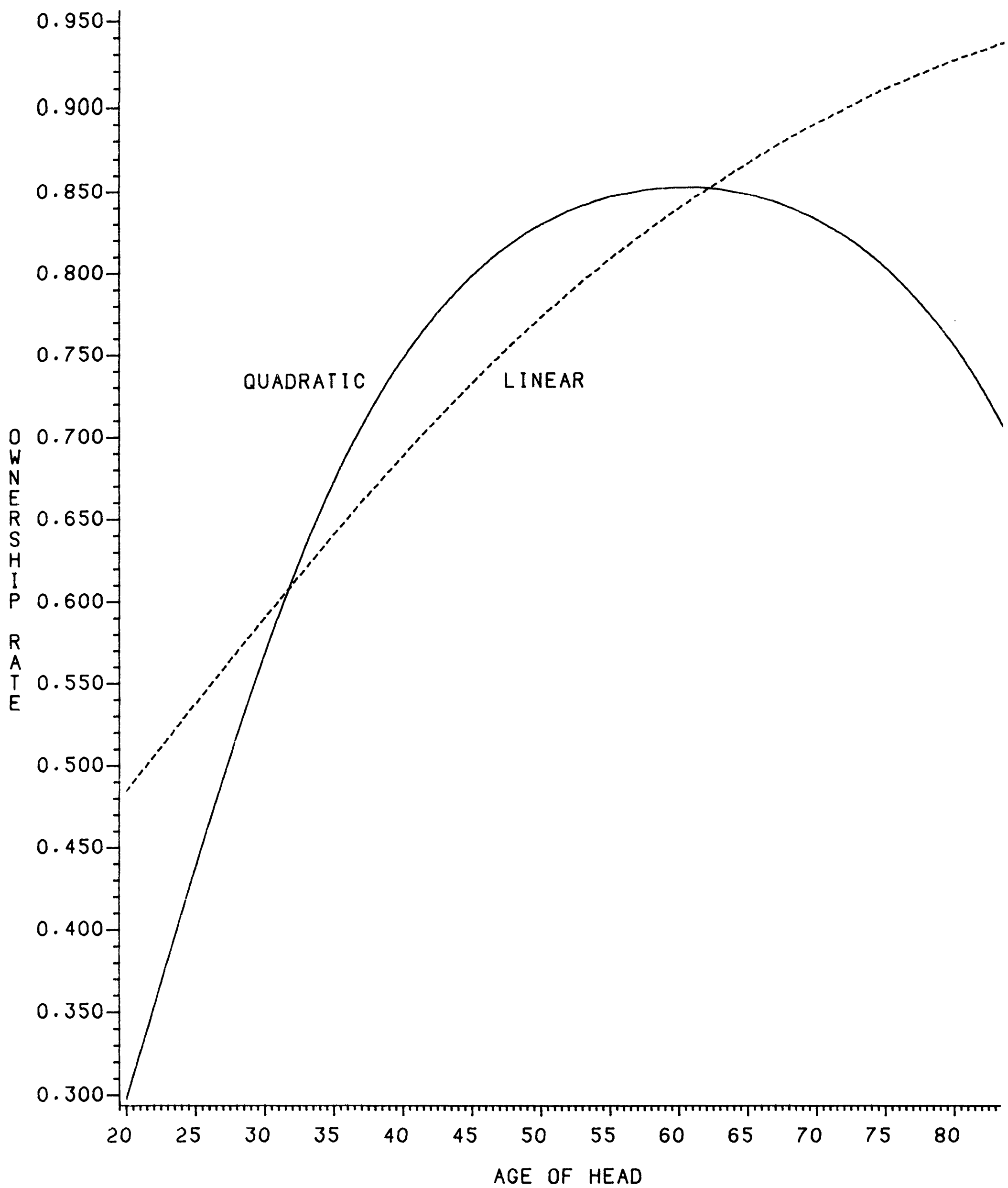


Table 2. OLS and Logit Analyses of Ownership as Functions of Age of Head and Age Squared (couples)

| | OLS | | Logit | |
|---|---|---|---|---|
| Variable | Coeff. | t | Coeff. | t |
| Age of Head | .04076 | 5.872 | .09662 | 5.27166 |
| Age Squared | .00034 | -4.856 | .00080 | -4.26851 |
| Constant | -.34763 | -2.195 | 2.95953 | 7.23240 |
| | ($R^2$ = .09327) | | ($R^2$ = .08003)* | |
| | Goodness of fit Chi square=65.0, d.f.=63, p=.405 | | | |

*Efron $R^2$ (Capps and Kramer, 1985. For other measures of goodness of fit for logit, see Aldrich and Nelson, 1984.)

Table 3. OLS and Logit Analyses of Ownership Preference as Functions of Age of Head and Age Squared (couples)

| | OLS | | Logit | |
|---|---|---|---|---|
| Variable | Coeff. | t | Coeff. | t |
| Age of Head | .01114 | 2.483 | .05176 | 2.05272 |
| Age Squared | -.00013 | -2.805 | -.00059 | -2.40704 |
| Constant | .70243 | 6.866 | 5.21684 | 8.83125 |
| | ($R^2$ = .01686) | | ($R^2$ = .00638) | |
| | Goodness of fit Chi square=60.96, d.f.=63, p=0.549 | | | |

0.66 for logit resulting in a difference of 0.06. Partly because of the low variance in ownership preference, neither the OLS nor the logit regressions "explain" much of the variance in the dependent variable.

The average ownership *gap* by age can be seen if actual ownership is graphed with ownership preference. The OLS comparisons are presented in Figure 4. The predicted ownership preference at age 20 is 0.88, while the predicted ownership rate is 0.33, so that the predicted ownership is 54 percentage points lower than the ownership preference. The gap steadily narrows, reaching a minimum at age 70 with a gap of only 0.03. With the OLS estimates, ownership preference is always greater than the ownership rate. The logit analyses reveal a different picture (Figure 5). At age 20, the predicted ownership rate is 59 percentage points lower than the ownership preference. By age 80, ownership preference has dropped to the level of the ownership rate. After age 80, slightly more people own than prefer to own.

**FIGURE 2 OLS AND LOGIT ANALYSES OF OWNERSHIP RATE AS FUNCTIONS OF AGE OF HEAD AND AGE SQUARED (COUPLES)**

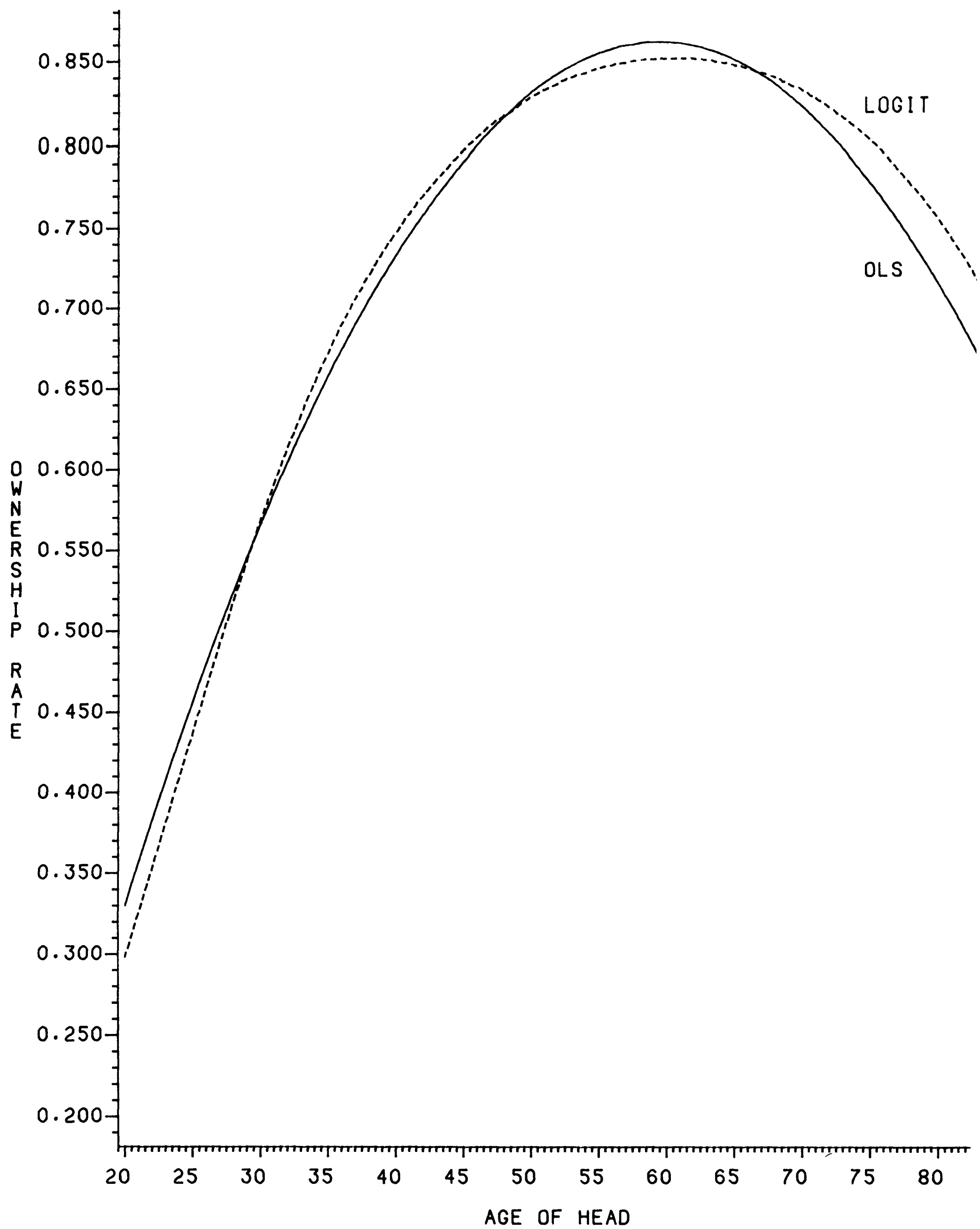


**FIGURE 3 OLS AND LOGIT ANALYSES OF OWNERSHIP PREFERENCE AS FUNCTIONS OF AGE OF HEAD AND AGE SQUARED (COUPLES)**

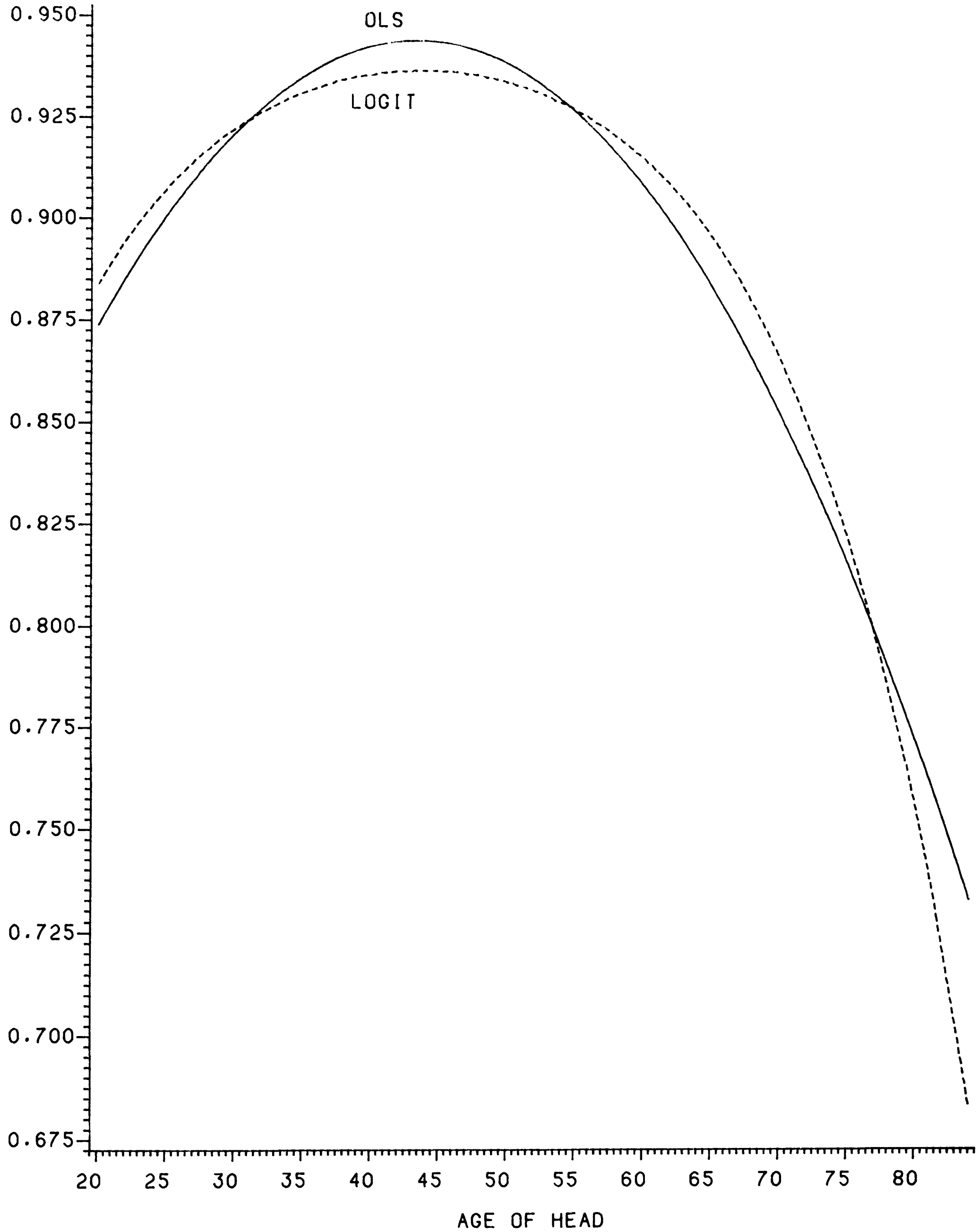


**FIGURE 4 OLS ANALYSIS OF OWNERSHIP AND OWNERSHIP PREFERENCE AS FUNCTIONS OF AGE OF HEAD AND AGE SQUARED (COUPLES)**

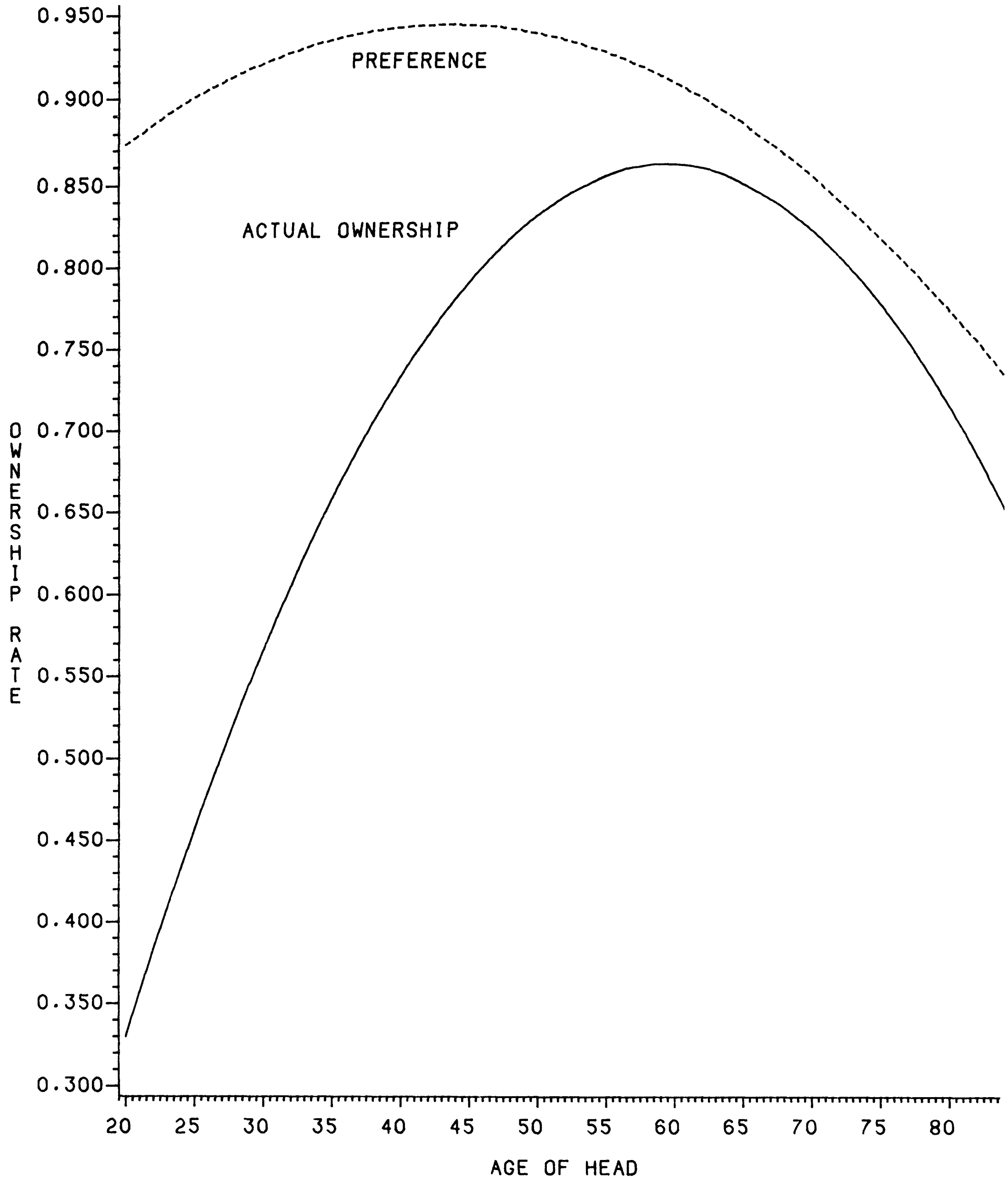
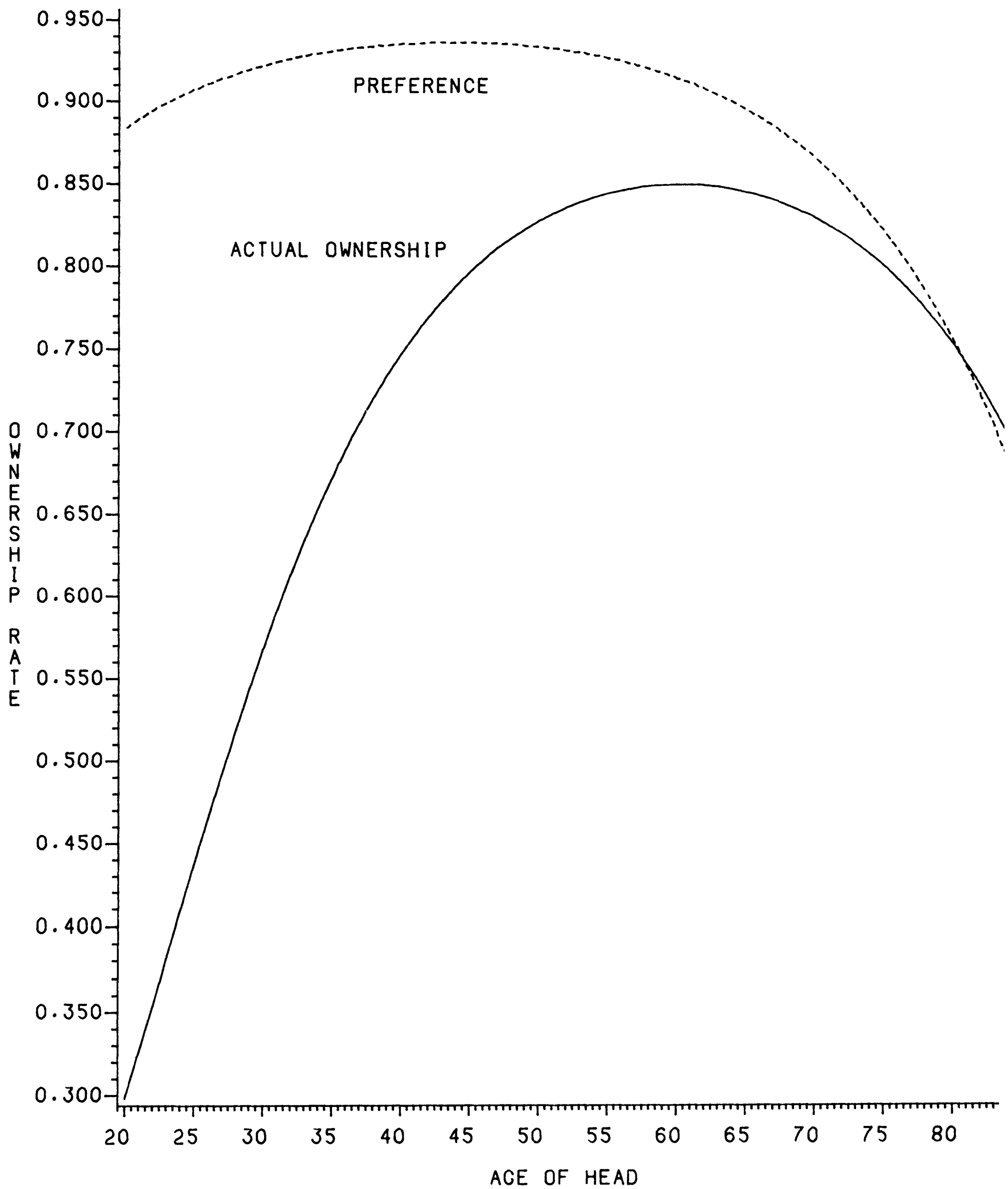


**FIGURE 5 LOGIT ANALYSES OF OWNERSHIP AND OWNERSHIP PREFERENCE AS FUNCTIONS OF AGE OF HEAD AND AGE SQUARED (COUPLES)**



Although clearly different conclusions might be drawn from Figures 4 and 5, one should not make too much of the differences. The important difference between the two graphs is at age 85, which is at the extreme value of the sample. The low t statistics for the age coefficients in the preference logit and OLS regressions mean that the predictions based on those coefficients are not very precise.

One should conclude that a narrowing of the gap takes place for the elderly as age increases, but one should not conclude that gap actually closes.

## *DISCUSSION*

Obviously, the appropriate procedure to use with dichotomous dependent variables is logit or some other procedure designed specifically for that purpose. OLS regression is never the *preferred* method of analysis of such dependent variables. However, for exploratory hypothesis testing, OLS is clearly to be preferred because of its ease of use and interpretation and lower computation cost. If the main variable of interest in a study has a small effect not close to significance in an OLS analysis with a dichotomous variable (Guy and Pohl, 1983), it is unlikely that logit would show a different result.

However, if a variable is of borderline significance or nonsignificance in an OLS analysis, or policy actions are to be taken based on the analysis, logit or probit analysis is advisable for comparison. If some subgroups in the sample have mean values of the dependent variable very close to 1.0 or to 0 (Goodman, 1975), OLS may yield predicted probabilities greater than 1 or less than 0. Logit or probit is advisable for such regressions, unless one is interested only in exploratory hypothesis testing. (An example of such a possibility is a regression of ownership preference on current ownership status, where virtually all nonelderly home owners may prefer ownership.) One should not, however, neglect measurement problems or model specification issues in the process.

## *REFERENCES*

Aigner, D.J., Goldberger, A.S., and Kalton, G. On the explanatory power of dummy variable regressions. *International Economic Review*, 1975, *16*, 503-510.

Aldrich, J.H., and Nelson, F.D. *Linear Probability, Logit, and Probit Models*. Beverly Hills: Sage Publications, 1984.

Burgess, S.L. Determinants of home ownership: A comparison of single female and single male headed households. *Housing and Society*, 1982, *9*, 87-94.

Capps, O. Jr., and Kramer, R.A. Analysis of food stamp participation using qualitative choice models. *American Journal of Agricultural Economics*, 1985, *67* (1), 49-59.

Chi, P.S. Determinants of changes in housing tenure status in the United States. *Housing and Society*, 1985, *11*, 121-135.

Derrick, F.W. Notes on the use of qualitative variables in regression. *Home Economics Research Journal*, 1979, *7*, 265-269.

Goodman, L.A. The relationship between the modified and more usual multiple regression approaches to the analysis of dichotomous variables. In D.R. Heise (Ed.), *Sociological Methodology*. San Francisco: Jossey-Bass, 1976, 83-110.

Guy, R.F., and Pol, L.G. Racial discrimination in home ownership: A reevaluation of the preference hypothesis. *Housing and Society*, 1983, *10*, 117-124.

Hanna, S., and Lindamood, S. Housing preferences of blacks and whites in Montgomery, Alabama. *Housing and Society*, 1979, *6*, 39-47.

Hrozencik, Diane F. Binomial logit analysis. *Proceedings of the American Council on Consumer Interests*, 1984, 147-150.

Kinsey, Jean, and Lane, Sylvia. Race, housing attributes and satisfaction with housing. *Housing and Society*, 1983, *10*, 98-116.

Maddala, G.S. *Limited-dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press, 1983.

Morris, E.W., and Winter, M. *Housing, Family and Society*. New York: John Wiley and Sons, 1978.

Morris, E.W., Winter, M., and Sward, M.A. Reporting error and single family home ownership norms and preferences. *Housing and Society*, 1984, *11*, 82-97.

Pindyck, R.S., and Rubinfeld, D.R. *Econometric Models and Forecasts*. New York: McGraw-Hill, 1976.

Pol, Louis G. Anticipated discrimination in the home lending market. *Housing and Society*, 1981, *8* (3), 3-11.

Winter, M., and Morris, E.W. Housing conditions, satisfaction and conventionality: An analysis of female headed households. *Housing and Society*, 1982, *9*, 70-86.