# ESTIMATING LINEAR PROBABILITY FUNCTIONS: A COMPARISON OF APPROACHES

David L. Debertin, Angelos Pagoulatos, and Eldon D. Smith

A linear probability function permits the estimation of the probability of the occurrence or non-occurrence of a discrete event. Nerlove and Press (p. 3-9) outline several statistical problems that arise if such a function is estimated via OLS. In particular, heteroskedasticity inherent in such a regression model leads to inefficient estimates of parameters (Amemiya 1973, Horn and Horn). Moreover, without restrictions on the conventional OLS model, probability estimates lying outside the unit (0-1) interval are possible (Nerlove and Press). Goldberger and Kmenta suggest two approaches for alleviating the heteroskedasticity problems inherent in the OLS regression model. Logit analysis will also alleviate heteroskedasticity problems and ensure that estimated probabilities will lie within the unit interval (Amemiya 1974, Hauck and Donner, Hill and Kau, Horn and Horn, Horn, Horn, and Duncan, Theil 1970).

We use data from a study conducted by Smith, Deaton, and Kelch (SDK) to compare the impacts of heteroskedasticity adjustments on the OLS model with those obtained from a model estimated by means of logit analysis. SDK estimated a linear probability function in an effort to isolate determinants of industrial location decisions.

Batie criticized the SDK study because adjustments were not made for the heteroskedasticity problem. The dependent variable used in the study was the dichotomous decision of one or more firms to locate in the community. Independent variables included 11 variables hypothesized to influence firm managers in making this decision: (1) a measure of labor availability, (2) a measure of industrial site quality, (3) the availability of financing, (4) ownership of the site, (5) the distance of the site to a standard metropolitan statistical area, (6) the dollars spent per pupil on education, (7) the presence or absence of a college, (8) the presence or absence of an interstate highway, (9) population of the community, (10) employment in manufacturing in the community, and (11) fire protection rating. A more detailed description of each variable along with a theoretical justification for the model was given by Smith, Deaton, and Kelch. The model was applied to cross-sectional data for 565 Kentucky and Tennessee communities.

## ADJUSTMENTS TO THE OLS MODEL

### The Heteroskedasticity Problem

A heteroskedasticity problem results in inefficient parameter estimates, leads to problems in interpreting statistical tests of significance for individual regression parameters, and casts doubt on statistics such as coefficients of determination and F-ratios for the entire equation. Estimation of an OLS function were the dependent variable is dichotomous (0 or 1) via OLS will violate the OLS assumption that

$$(1) \quad E(\varepsilon\varepsilon') = \sigma^2 I$$

where

$\varepsilon$ = an n × 1 vector of residuals, where n = the number of observations

$\sigma^2 I$ = an n × n matrix whose diagonal elements are equal to $\sigma^2$ and whose off-diagonal elements are zero.

We designate the observations from the $t^{th}$ row of X as $X_t'$ of dimension 1 × p, t = 1, 2, ..., n.

Then for any $t^{th}$ observation, when $Y_t$ assumes a value of zero, $\varepsilon_t$ is equal to $-X_t'\beta$. $\beta$ is a parameter vector of dimension p × 1. The variance-covariance matrix of the disturbances is a heteroskedastic matrix whose diagonal elements are equal to $X_t'\beta(1 - X_t'\beta)$ and whose off-diagonal elements are zero (Johnston, p. 227). When $Y_t$ assumes a value of one, $\varepsilon_t = 1 - X_t'\beta$. Because $\varepsilon_t$ assumes the same distribution as $Y_t$, the assumption of normality is untenable.

### The Goldberger Procedure

Goldberger and Gujariti both provide a simple procedure to adjust for the heteroskedasticity problem. They suggest first estimating the linear probability function via OLS to obtain probability estimates from

$$(2) \quad \widehat{P}_t = X_t'b$$

where b is a p × 1 parameter vector estimator.

Estimates of the diagonal elements of the residual variance-covariance matrix for each observation become

David L. Debertin is Professor, Angelos Pagoulatos is Associate Professor, and Eldon Smith is Professor, Department of Agricultural Economics, University of Kentucky.

(3) $w_t = \widehat{P}_t(1 - \widehat{P}_t)$.

As a second step, all variables are divided by

(4) $w_t^{.5} = [P_t(-P_t)]^{.5}$

and the equation is re-estimated. The procedure is a feasible Aitkens estimator where $\widehat{\Omega}$ is substituted for $\Omega$.

(5) $b^* = [X'\Omega^{-1}X]^{-1} Y$

b* and Y are vectors, not scalars.

In this case, $\Omega^{-1}$ is a diagonal matrix where $1/\widehat{P}_t[1 - \widehat{P}_t)$ are the diagonal elements. However, nothing is built into the OLS procedure conducted in step 1 to ensure that $0 < \widehat{P}_t < 1$. If $\widehat{P}_t \geqslant 1$ or $\widehat{P}_t \leqslant 0$, then $\widehat{P}_t(1 - \widehat{P}_t) \leqslant 0$, and the data transformation breaks down. Unfortunately, linear probability functions estimated from typical data sets often generate $\widehat{P}_t$ that are $\geqslant 1$ or $\leqslant 0$.

In the SDK study, 55 of 565 or nearly 10 percent of the observations generated predicted values outside (0, 1). We used several approaches to remedy the problem. The first approach was to delete observations for which the probabilities estimated from the OLS equation were outside (0, 1). Data were transformed for the 510 remaining observations via the Goldberger procedure and the equation was re-estimated. Results suggested that despite the removal of information on observations whose values for independent variables were far from the sample means, the $R^2$ did not change substantially and the F-ratios decreased only slightly (Table 1). Impact appeared to be greater on coefficients of discrete independent variables than on those of continuous independent variables. A number of variables significant at the .05 level in the original model became nonsignificant at the chosen probability level.

The second approach was to re-estimate the entire question, and in those instances where $\widehat{P}_t(1 - \widehat{P}_t) \leqslant 0$, $\widehat{P}_t(1 - \widehat{P}_t)$ was set equal to some small positive number. The choice of the small positive value is arbitrary, and as we show, results depend on selected values. This is a key problem with the approach.

The value .00001 was the first small positive number used. For those observations where $\widehat{P}_t(1 - \widehat{P}_t)$ is a very small positive number, the transformed dependent variable as well as independent variables become extremely large. The result is an enormous spurious increase in the coefficient of determination for the transformed regression equation (Table 1). The $R^2$ increased from 36 to 81 percent. Coefficients for the transformed data were in several instances substantially different from the OLS results. Coefficients that were significant in the OLS results became nonsignificant. Certain coefficients that were nonsignificant became significant. Results were very unstable.

Coefficients were then derived by setting $\widehat{P}_t(1 - \widehat{P}_t)$ equal to .01 for those observations in which $\widehat{P}_t(1 - \widehat{P}_t) \leqslant 0$, transforming the data, and re-estimating the equation. The $R^2$ of 64 percent, though higher than the 36 percent for the OLS equation, was less than the 81 percent obtained when .0001 was used. This outcome was expected, because the impact of the adjustment on the revised variables was

TABLE 1. REGRESSION STATISTICS FOR O L S LINEAR PROBABILITY MODELS AND MODELS ADJUSTED FOR HETEROSKEDASTICITY

| | OLS Regression | | Corrected $\hat{P}_t(1-\hat{P}_t) \leq 0$ deleted | | Corrected $\hat{P}_t(1-\hat{P}_t) \leq 0 =$ .00001 | | Corrected $\hat{P}_t(1-\hat{P}_t) \leq 0 =$ .01 | | Corrected $\hat{P}_t(1-\hat{P}_t) \leq 0 =$ .1 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Independent Variable | Coefficient | Standard Error | Coefficient | Standard Error | Coefficient | Standard Error | Coefficient | Standard Error | Coefficient | Standard Error |
| Intercept | -0.224897 | .099336 | +0.482187 | .149992 | -0.005905 | .064621 | -0.112122 | .063465 | -0.228815 | .087612 |
| Labor Availability | +0.001060 | .001038 | -0.000713 | .000846 | -0.000703 | .000519 | +0.000256 | .000702 | +0.000878 | .000949 |
| Site Quality | +0.004013 | .000944 | +0.002418 | .000976 | +0.024863 | .009893 | +0.003279 | .000989 | +0.003444 | .000929 |
| Financing | +0.196137 | .037380 | +0.121825 | .048926 | +0.335243 | .098457 | +0.232189 | .041382 | +0.195553 | .037903 |
| Site Ownership | +0.145299 | .043847 | +0.187650 | .049232 | +0.099776 | .077509 | +0.155146 | .050032 | +0.162232 | .046141 |
| Miles to SMSA | -0.000105 | .000491 | -0.000110 | .000371 | +0.000165 | .000229 | -0.000228 | .000274 | -0.000193 | .000424 |
| Dollars Education Expenditure per Pupil | +0.000400 | .000178 | -0.000114 | .000092 | +0.000002 | .000145 | +0.000262 | .000128 | +0.000456 | .000159 |
| College Present or absent | +0.139070 | .078893 | +0.283059 | .081455 | -1.533711 | .854180 | +0.305049 | .076978 | +0.168094 | .076351 |
| Interstate Highway | +0.064513 | .032587 | -0.002510 | .019861 | -0.024100 | .031201 | +0.011392 | .021589 | +0.037978 | .028728 |
| Community Population | +0.000000 | .000007 | -0.000002 | .000007 | -0.000005 | .000003 | -0.000000 | .000004 | -0.000000 | .000006 |
| Manufacturing Employment | +0.000005 | .000016 | +0.000035 | .000015 | -0.000043 | .000008 | -0.000032 | .000009 | -0.000005 | .000013 |
| Fire Protection rating | +0.033860 | .014221 | -0.003769 | .012591 | +0.034889 | .031831 | +0.030932 | .011587 | +0.035092 | .012522 |
| $R^2$ | 0.3696 | | 0.3505 | | 0.8113 | | 0.6364 | | .5524 | |
| F | 29.48 | | 24.43 | | 198.11 | | 80.65 | | 56.87 | |
| N | 565 | | 510 | | 565 | | 565 | | 565 | |

lessened. Coefficients tended to be closer to the OLS results, with some differences. For example, the coefficient on manufacturing employment was nonsignificant in the OLS results, significant at the .05 level and positive when .0001 was used, but significant at the .05 level and *negative* when .01 was used.

Finally, the equation was re-estimated with .1 for the value of $\widehat{P}_t(1 - \widehat{P}_t)$ when $\widehat{P}_t$ was outside the (0, 1) range. The $R^2$ was .55 and the regression coefficients, with some exceptions, were similar to the OLS results. Manufacturing employment became nonsignificant, but this time with a negative sign.

### Kmenta's Iterative Procedure

Kmenta argues that the problems of heteroskedasticity can be resolved better through an iterative application of generalized least squares (p. 265). He suggests, as does Goldberger, that an estimate of the residual variance-covariance matrix can be obtained from the first pass of the OLS. By use of the GLS procedure, an "improved" estimate of the residual variance-covariance matrix is obtained, and the procedure is used again to re-estimate the regression parameters. The procedure can be applied as many times as desired. One of the GLS assumptions is that $\Omega$ is positive definite (Johnston). Though negative values of $\widehat{P}_t(1 - \widehat{P}_t)$ could appear on the diagonal of $\Omega$ in violation of this assumption, results using a negative variance would be nonsensical. A positive variance can be calculated only if more than one value for each X observation is available. This is not feasible with usual economic data. Otherwise, values for observations in which $\widehat{P}_t(1 - \widehat{P}_t) \leq 0$ must be made positive or these observations must be deleted.

The Kmenta procedure was applied with the assumption that values of $\widehat{P}_t(1 - \widehat{P}_t) \leq 0$ were equal to .01. The choice was arbitrary. Covariances of residuals were restricted to zero. The results, summarized for four iterations in Table 2, provide little evidence to support the notion that the iterative procedure will ultimately provide "better" estimates of regression parameters and their associated standard errors. Coefficients tend to be rather unstable through successive iterations. Coefficients of determination and F-ratios actually decrease slightly. In short, the iterative procedure when applied to real data does not appear to provide results consistent with the theoretical promise inherent in the GLS estimation technique.

## LOGIT ANALYSIS

Probit, logit, and tobit analysis have been proposed as techniques for ensuring that predicted probabilities always lie between zero and one (Penn, Witherington and Wills). Predicted values initially obtained from the sample data are transformed via a normal cumulative function that can be represented by a sigmoid curve (McFadden 1974, Nerlove and Press, Theil 1971). The logit transformation has been the most popular among researchers working with economic data. The logit transformation defines the probability of an event occurring as:

(6) $P_t = 1/[1 + \exp(X_t'b)]$.

TABLE 2. REGRESSION STATISTICS FOR OLS LINEAR PROBABILITY MODELS HETEROSKEDASTICITY ADJUSTMENTS VIA ITERATIVE GLS PROCEDURES

| | Iteration 1 | | Iteration 2 | | Iteration 3 | | Iteration 4 | |
|---|---|---|---|---|---|---|---|---|
| | Coefficient | Standard Error | Coefficient | Standard Error | Coefficient | Standard Error | Coefficient | Standard Error |
| Intercept | -.112212 | .063465 | +.003085 | .064197 | -.173233 | .069937 | -.013557 | .064014 |
| Labor Availability | +.000256 | .000702 | -.000417 | .000637 | +.000422 | .000678 | -.000314 | .000647 |
| Site Quality | +.003279 | .000989 | +.003702 | .001025 | +.002789 | .001015 | +.003463 | .001009 |
| Financing | +.232189 | .041382 | +.212023 | .043514 | +.194280 | .043843 | +.203548 | .041979 |
| Site Ownership | +.155146 | .050032 | +.170318 | .053264 | +.228315 | .048875 | +.173358 | .054011 |
| Miles to SMSA | -.000228 | .000274 | -.000186 | .000276 | -.000086 | .000289 | -.000032 | .000265 |
| Dollars Education Expenditure per pupil | +.000262 | .000128 | +.000036 | .000129 | +.000341 | .000119 | +.000022 | .000132 |
| College Present or Absent | +.305049 | .076978 | +.120388 | .070086 | +.270637 | .079666 | +.112950 | .072745 |
| Interstate Highway | +.011392 | .021589 | -.008199 | .019681 | +.022154 | .021686 | -.006182 | .019239 |
| Community Population | -.000000 | .000004 | +.000007 | .000005 | -.000009 | .000004 | +.000005 | .000006 |
| Manufacturing Employment | -.000032 | .000009 | -.000008 | .000010 | -.000008 | .000009 | -.000000 | .000010 |
| Fire Protection rating | +.030932 | .011587 | +.020013 | .011254 | +.042211 | .011589 | +.025024 | .012234 |
| $R^2$ | .63637 | | .64240 | | .63766 | | .61444 | |
| F | 80.65 | | 82.78 | | 81.10 | | 73.44 | |
| N | 565 | | 565 | | 565 | | 565 | |

Hence, $X_t'b$ need only lie between $-\infty$ and $+\infty$ for the probability to lie between zero and one.

Estimation of the logit function is straightforward for controlled experiments in which cell treatments are replicated, because probability estimates of the occurrence of an event for each cell are easily calculated. Economic data from uncontrolled experiments pose greater problems.

In our analysis, we adopt the logit method suggested by Berkson, refined by Theil (1970), and used by Li which employs categorization of variables and cell frequency counts to derive the logit estimate. Nerlove and Press outline a procedure for obtaining maximum likelihood estimates of probability function parameters from economic data without relying on cell frequency counts. However, the requisite program requiring iterative gradient maximization of the likelihood function is computationally burdensome and not readily available. Moreover, McFadden has shown that the cell frequency count method is not only preferable on computational grounds, but also is asymptotically equivalent to the maximum likelihood procedure. Small sample properties of either approach are the subject of great controversy among researchers using logit and probit analysis (Eeron, Sanathanan, Witherington and Wills).

The logit function we estimated is:

$$(7) \quad \log[\tilde{P}_t/(1 - \tilde{P}_t)] = X_t'b + e_t$$

where

> $\tilde{P}_t$ is the estimated probability of the firm locating in the community.

To obtain $\tilde{P}_t$ estimates, we categorized observations in the data set into groups of six based on similarities in the magnitude of each independent variable. A probability was assigned to each observation in the group on the basis of the proportion of the total observations in which an industrial plant was established. The equation was then estimated via least squares. To obtain the probabilities from the logit function contained within the unit interval, $X_t'b$ estimated from equation 7 was inserted into equation 6.

Figure 1 illustrates how the calculated probabilities obtained from the logit function differ from those obtained from the simple OLS linear probability function. Even though the simple correlation between the two probability estimates is .94, the estimates differ markedly at the extremes of the distribution. Data charted in Figure 1 closely correspond to the results suggested by the figure in the Nerlove and Press report (p. 4).



Table 3 summarizes the results of the logit function estimation. Results from the logit function estimation are superior to those obtained from the OLS methods using the (0-1) dummy as the dependent variable, even when adjustments are made for the heteroskedasticity problem. Coefficients for most variables are substantially larger in relation to the respective standard errors for the logit model

FIGURE 1. RELATIONSHIP BETWEEN ESTIMATED X b FROM LINEAR PROBABILITY FUNCTION AND THE LOGIT FUNCTION PROBABILITY

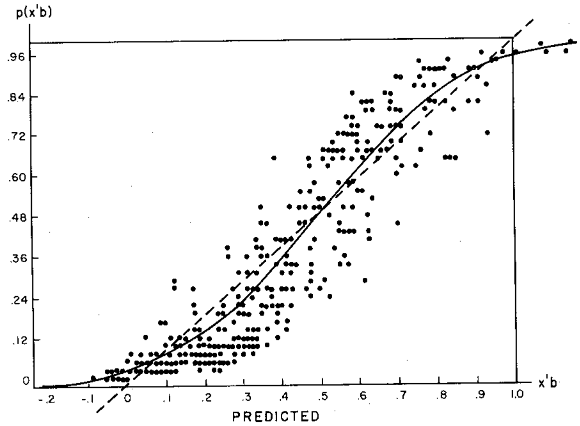


TABLE 3. ESTIMATES OF LOGIT MODEL PARAMETERS

Table 3. Estimates of Logit Model Parameters

| Independent Variable | Coefficient | Standard Error |
|---|---|---|
| Intercept | -7.255982 | .422906 |
| Labor Availability | +.017765 | .004419 |
| Site Quality | +.015468 | .004017 |
| Financing | +1.314700 | .159138 |
| Site Ownership | +1.413099 | .186669 |
| Miles to SMSA | +.007963 | .002091 |
| Dollars Educational Expenditure per Pupil | +.003801 | .000756 |
| College Present or Absent | +.747906 | .335876 |
| Interstate Highway | +.249169 | .138734 |
| Community Population | -.000029 | .000029 |
| Manufacturing Employment | +.000037 | .000068 |
| Fire Protection Rating | +.608331 | .060543 |

$R^2 = .6529$
$F = 94.583$
$N = 565$

than for the unadjusted or adjusted OLS models. For example, the coefficient on labor availability is several times its standard error in the logit model, but at best only slightly larger than its standard error for any of the other models. A similar result is obtained for several other variables, including site quality, financing, educational expenditures, and fire protection ratings. The logit function suggests that variables under control of the local decision makers are more systematically related on the plant location decision than has previously been revealed in the simple OLS results. A comparison of parameters from the approaches reveals that signs on variables remain relatively stable. Miles to SMSA is the only variable for which the sign changed when the methods were compared.

## SUMMARY

In estimating the probability of an occurrence of a discrete event, the particular method of model estimation is extremely important. Simple adjustments to alleviate the heteroskedasticity problem did little to improve the results and introduced new problems in the estimation. The logit function supplied substantially better results than the OLS function, even when adjustments for heteroskedasticity were made. The improvements are indicated by a substantial increase in the magnitude of most coefficients in relation to the respective standard errors, as well as increases in the usual measures of explained variation, such as $R^2$ and the F-ratio. As a result of the logit function estimation using the SDK data, we now believe that community decision makers can have much greater impact on the industrial location decision than was previously believed. Variable under the control of decision makers include site quality and ownership, financing, educational expenditures, and fire protection ratings. Though the OLS estimates have the convenience of being directly interpretable, the logit model will be more reliable for establishing the effectiveness of community actions.

## REFERENCES


Amemiya, Takeshi. "Bivariate Probit Analysis, Minimum Chi Square Methods." *J. Amer. Statist. Assoc.* 69(1974):940-4.

———. "Regression Analysis When the Dependent Variable is Truncated Normal." *Econometrica* 41(1973):997-1016.

Batie, Sandra. "Discussion: Location Determinants of Manufacturing Industry in Rural Areas." *S. J. Agr. Econ.* 10(1978):33-7.

Berkson, G. "Maximum Liklihood and Minimum Chi Square Estimates of the Logistic Function." *M. Amer. Statist. Assoc.* 50(1955):130-61.

Eeron, Bradley. "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis." *J. Amer. Statist. Assoc.* 70(1975):892-8.

Goldberger, Arthur. *Econometric Theory.* New York: John Wiley & Sons, Inc., 1964.

Gujariti, Damodar. *Basic Econometrics.* New York: McGraw Hill Book Company, 1978.

Hauck, Walter W., Jr. and Allan Donner. "Wald's Test as Applied to Hypothesis on Logit Analysis." *J. Amer. Statist. Assoc.* 72(1977):851-3.

Hill, Lowell, and Paul Kau. "Application of Multi-variate Probit to a Threshold Model of Grain Dryer Purchasing Decisions." Amer. J. Agr. Econ. 55(1973):19-27.

Horn, Susan and Roger A. Horn. "Comparison of Estimators of Heteroskedastic Variances in Linear Models." *J. Amer. Statist. Assoc.* 70(1975):872-9.

Horn, Susan D., Roger A. Horn, and David B. Duncan. "Estimating Heteroskedastic Variances in Linear Models." *J. Amer. Statist. Assoc.* 70(1975):380-5.

Johnston, G. *Econometric Methods.* New York: McGraw-Hill Book Company, 1972.

Kmenta, Jan. *Elements of Econometrics.* New York: Macmillan Company, 1971.

Li, Mingche M. "A Logit Model of Homeownership." *Econometrica* 45(1977):1081-95.

McFadden, Daniel. "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics,* Paul Zarembka, editor. New York: Academic Press, 1974, 105-42.

———. "Quantal Choice Analysis: A Survey." *Ann. Econ. and Soc. Meas.* 5(1976):363-90.

Nerlove, Marc and S. James Press. "Univariate and Multivariate Log Linear and Logistic Models." Rand Corp. Rept. R-1306-EDA/NIH 1973.

Penn, J. B. "On Probits, Logits and Tobits: A Description and Discussion of Applications to Economics." Purdue University, 1971.

Sanathanan, Lilitha. "Some Properties of the Logistic Model for Dichotomous Response." *J. Amer. Statist. Assoc.* 69(1974):744-9.

Smith, Eldon, Brady Deaton, and David Kelch. "Location Determinants of Manufacturing Industry in Rural Areas." *S. J. Agr. Econ.* 10(1978):23-32.

Theil, Henri. "On the Estimation of Relationships Involving Qualitative Variables." *Amer. J. Soc.* 76(1970):103-54.

———. *Principles of Econometrics.* New York: John Wiley & Sons, Inc., 1971.

Witherington, Moffat Patrick and Cleave E. Wills. "The Dichotomous Dependent Variable: A Comparison of Probit Analysis and Ordinary Least Squares Procedures by Monte Carlo Analysis." *Res. Bull. 657,* University of Massachusetts, College of Food and Natural Resources, 1978.