COWLES FOUNDATION DISCUSSION PAPER NO. 1

The Application of Multivariate Probit Analysis to Economic Survey Data

James Tobin

July 14, 1955

(As revised December 1, 1955)

The Application of Multivariate Probit Analysis to Economic Survey Data

Analysis of economic surveys of samples of households often has the objective of estimating the relationship of a dependent variable to a set of independent variables and of testing hypotheses about that relationship. Typically the dependent variable is a measure reflecting some kind of household behavior or decision, while the independent variables represent characteristics over which the household has less control, at least in the shortrun. For example, explanation of the variation among households in annual food expenditure may be sought in such independent variables as family income, family size, occupation of head of household, age of head of household, and location. In this example and in many other cases, the dependent variable can take on a large number of possible values along a natural scale. Thus food expenditure, if households report to the nearest dollar, can in principle be any nonnegative integer, though its realistic range is doubtless limited. For dependent variables of this kind, the theory of multiple regression -- including analysis of variance and covariance -- provides an appropriate statistical model.

Sometimes, however, the dependent variable of interest is dichotomous. It can take on only two values, which can for convenience be designated as 1 and 0. The household either owns a house or does not own one; the household either bought a new car last year or did not buy one; or, to cite a variable from a neighboring social science, the head of the household either likes Ike or does not. As in the food expenditure example, a variety of independent variables may be associated with the differences between home-owners and non-home-owners, or car-buyers and non-buyers, or supporters and opponents of a Presidential candidate. But the association is necessarily of a different kind. An increase in income may be expected to result in an increase in the food expenditure of a given household. An increase in income may also turn a household from a non-owner into a home-owner. But it cannot make a home-owning household into any more of a home-owner than the household already is.

In the case of food expenditure, it is important to know the exact level of the household's income. In the case of ownership, the important thing is whether or not this income exceeds some critical value.

Multiple regression is accordingly not an appropriate model for a dichotomous dependent variable. By the definition such a variable, its expected value must always be in the interval (0,1), whatever the value of the independent variables. This condition cannot be maintained if the expected value is assumed, as in multiple regression, to be a linear combination of the independent variables. Moreover, the multiple regression model assumes, inappropriately for this case, that the distribution of the dependent variable around its expected value is independent of the level of that expected value. For a dichotomous variable, an expected value of .8 means a probability of .8 that the value will deviate from expectation by +.2 and a probability of .2 that the deviation will be -.8, while an expected value of .4 means deviations of +.6 with probability .4 and deviations of -.4 with probability .6.

Probit analysis (see Finney [8]) provides an appropriate model. In biological assay, probit analysis is used to determine the relationship between the probability that organisms will be killed to the strength of the dose of poison administered to them. The dependent variable, for each organism in the sample, is dichotomous: killed or not killed. Each organism is assumed to have a dosage threshold, such that a stronger dose will kill that organism and a weaker dose will not. Over the population of organisms of a given kind, the logarithms of these dosage thresholds are assumed to be normally distributed, with mean and standard deviation estimated from the data by maximum likelihood. The analogous use of probit analysis in economic surveys is illustrated by its application by Farrell [7] to the relationship between ownership of automobiles and income. In Farrell's application, the dependent variable is defined by whether or not the household owned a car of a given age or younger. Each household is assumed to have an income threshold, such that if its income is bigger than the critical value it owns, while if its income is below the threshold it does not. The logarithms of the income thresholds are assumed to be normally distributed. The parameters of the distribution are estimated, by maximum likelihood from data giving the number of sample households observed to own and not to own at various income levels.*

---

* A different economic application of probit analysis, to a case where the dependent variable is multi-valued and naturally scaled, has been made by Aitchison and Brown, [1] and [2].

In Farrell's application there is only one independent variable, income, to which the probability of car ownership is related. But he and other econometricians are keenly aware that observed differences among households in sample surveys are attributable to a multiplicity of factors. It is not possible to duplicate the experimental control that is feasible in biological assay. Consequently, multivariate probit analysis, like its counterpart, multiple regression, is an essential tool for the analyst of economic surveys. Finney ([8], Chapter 7) explains and illustrates the extension of the Bliss-Fisher maximum likelihood solution to cases with two or more independent variables. The exposition of multivariate probit analysis which follows in the present paper is not fundamentally different from Finney's treatment. It is, however, oriented to the problems of economic surveys rather than those of dosage-mortality experiments. It differs also in using the exact maximum likelihood equations of Garwood [9] rather than the approximations of the Bliss-Fisher procedure used by Finney. Fewer iterations are required to compute the solutions of the Garwood equations, and the publication of new tables by Cornfield and Mantel [6] has removed the practical obstacles to the use of these equations.

The Model

Suppose that there is an index I, which is a linear combination of the various independent variables $X_1, X_2, \ldots X_m$ that determine whether the dependent variable W has the value 0 or 1 for a household.

(1) $$I = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

The assumption that I is a linear combination of the X's is neither more nor less restrictive than the similar assumption in multiple regression. There are various devices by which a linear combination of $(X_1, X_2, \ldots X_m)$ can represent a non-linear function of the observed variables. An X in the index may be the logarithm or the square, or some other function of one of the original observed variables. Or an X in the index may be the product, or some other function, of two or more of the other X's, in order to test and to estimate interactions as well as main effects.

Let I_i be the actual value of the index for the ith household, determined by evaluating (1) for the values of the independent variables that obtain for the ith household. Let $\bar{I}_i$ be the critical value of the index for the ith household: If the actual value of the index

I_i equals or exceeds the critical value $\bar{I}_i$, then W_i will be 1; if I_i is less than $\bar{I}_i$, then W_i will be 0.

(2) $$W_i = 1 \quad \text{for} \quad I_i \geq \bar{I}_i$$
$$W_i = 0 \quad \text{for} \quad I_i < \bar{I}_i$$

Over the population of households the critical values $\bar{I}_i$ are assumed to be normally distributed with mean 5* and standard deviation 1.

* This convention is usual in probit analysis, and the tables are set up accordingly.

This distribution reflects random differences among households, for example differences in personality and taste, that are not represented by any of the variables in the index. Some households would not own a new car unless their income was very high, while others require only a bare margin above subsistence levels to put them over the new-car threshold.

For a given value of the index, I, W will be equal to 1 for those individuals for whom $\bar{I}_i \leq I$ and W will be equal to 0 for those whose $\bar{I}_i > I$. The probability that, given I, W_i will be equal to 1 is therefore:

(3) $$\Pr(W = 1|I) = \Pr(\bar{I}_i \leq I) = P(I) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I-5} e^{-\frac{u^2}{2}} du .$$

Similarly, the probability that, given I, W will be equal to 0 is:

(4) $$\Pr(W = 0|I) = \Pr(\bar{I}_i > I) = 1 - P(I) = Q(I) = \frac{1}{\sqrt{2\pi}} \int_{I-5}^{\infty} e^{-\frac{u^2}{2}} du$$

The Maximum Likelihood Solution*

* The exposition of the maximum likelihood solution that follows is a mathematically simple extension of the Garwood solution ([9] and summary in [6]) to $m + 1$ dimensions.

A sample of observations at s distinct points $(X_{1j}, X_{2j}, \dots X_{mj})$ where $(j = 1, 2, \dots s)$ may be summarized as follows: Let n_j be the total number of observations at the j-th point. Let r_j be the number of those observations for which W was observed to be 1, and $n_j - r_j$ the number for which W was observed to be 0. The likelihood of the sample is a function of the values $(b_0, b_1, \dots b_m)$ assumed for the population parameters $(\beta_0, \beta_1, \dots \beta_m)$:

$$(5)\quad L(b_0, b_1, \dots, b_m) = \prod_{j=1}^{s} [P(b_0 + b_1X_{1j} + \dots + b_mX_{mj})]^{r_j} [Q(b_0 + b_1X_{1j} + \dots + b_mX_{mj})]^{n_j - r_j} .$$

Here, as in (3) and (4) $P(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x-5} e^{-\frac{u^2}{2}} du$ and $Q(x) = 1 - P(x)$.

Let $Y_j = b_0 + b_1X_{1j} + \dots + b_mX_{mj}$; $P_j = P(Y_j)$; $Q_j = Q(Y_j)$.

To find the maximum likelihood estimates of the population parameters, it is convenient to find values of the b's to maximize log L rather than L.

$$(6)\quad L^*(b_0,b_1, \dots, b_m) = \log L(b_0, b_1, \dots b_m) + \sum_{j=1}^{s} (r_j \log P_j + (n_j-r_j) \log Q_j)$$

The conditions for the maximum are the $m + 1$ equations determined by setting the partial derivatives of L^* equal to zero. Let $L_i^*(b_0, b_1, \dots b_m) = \frac{\partial L^*}{\partial b_i}$.

Let $Z(x) = \frac{dP(x)}{dx} = \frac{-dQ(x)}{dx} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and let $Z_j = Z(Y_j)$. Note that $\frac{dZ_j}{dY_j} = -Y_j Z_j$

Let X_0 be identically 1. The equations are:

$$(7)\quad L_i^*(b_0, b_1, \dots b_m) = \sum_{j=1}^{s}\left(r_j \frac{X_{ij}Z_j}{P_j} - (n_j - r_j)\frac{X_{ij}Z_j}{Q_j}\right) = 0 \qquad (i = 0, 1, 2, \dots m)$$

These non-linear equations can be solved by an iterative process. Let $(b_{00}, b_{10}, \dots b_{m0})$ be trial solutions. New estimates $(b_{00} + \Delta b_0, b_{10} + \Delta b_1, \dots b_{m0} + \Delta b_m)$ can be found by solving the following set of $m + 1$ linear equations, in which all the L_i^* are assumed to be linear between the trial solution and the real solution.

$$(8)\quad L_i^*(b_{00} + \Delta b_0, b_{10} + \Delta b_1, \dots b_{m0} + \Delta b_m) =$$

$$L_i^*(b_{00}, b_{10}, \dots b_{m0}) + \sum_{k=0}^{m} \Delta b_k L_{ik}^*(b_{00}, b_{10}, \dots b_{m0}) = 0 \qquad (i = 0, 1, 2, \dots m)$$

The second order derivatives L_{ik}^* are given by differentiating (7):

$$(9)\quad L_{ik}^*(b_0, b_1, \dots b_m) =$$

$$\sum_{j=1}^{s}\left(\frac{r_j X_{ij}X_{kj}(-P_jY_jZ_j - Z_j^2)}{P_j^2} - \frac{(n_j - r_j)X_{ij}X_{kj}(-Q_jY_jZ_j + Z_j^2)}{Q_j^2}\right)$$

$$(i, k = 0, 1, 2 \dots m)$$

Following the notation of Cornfield and Mantel [6], let:

$$w'_{j\,max} = \frac{Y_j Z_j}{P_j} + \frac{Z_j^2}{P_j^2} . \qquad w'_{j\,min} = -\frac{Y_j Z_j}{Q_j} + \frac{Z_j^2}{Q_j^2} .$$

(10) $$n_j w'_j = r_j w'_{j\,max} + (n_j - r_j) w'_{j\,min}$$

Also, let:

$$\Delta_{j\,max} = \frac{Z_j}{P_j} \qquad \Delta_{j\,min} = \frac{Z_j}{Q_j}$$

(11) $$n_j \Delta_j = r_j \Delta_{j\,max} - (n_j - r_j)\Delta_{j\,min}$$

Equations (7) can then be rewritten:

(12) $$L_i^*(b_0, b_1, \ldots b_m) = \sum_{j=1}^{s} X_{ij} n_j \Delta_j = 0 \qquad (i = 0, 1, 2, \ldots m)$$

If w'_j and Δ_j are evaluated for the trial-solution values of the b's, equations (8) can be rewritten:

(13)

$$\Delta b_0 \sum_{j=1}^{s} n_j w'_j + \Delta b_1 \sum_{j=1}^{s} X_{1j} n_j w'_j + \ldots + \Delta b_m \sum_{j=1}^{s} X_{mj} n_j w'_j = \sum_{j=1}^{s} n_j \Delta_j$$

$$\Delta b_0 \sum_{j=1}^{s} X_{1j} n_j w'_j + \Delta b_1 \sum_{j=1}^{s} X_{1j}^2 n_j w'_j + \ldots + \Delta b_m \sum_{j=1}^{s} X_{1j} X_{mj} n_j w'_j = \sum_{j=1}^{s} X_{1j} n_j \Delta_j$$

. . . .

. . . .

$$\Delta b_0 \sum_{j=1}^{s} X_{mj} n_j w'_j + \Delta b_1 \sum_{j=1}^{s} X_{mj} X_{1j} n_j w'_j + \ldots + \Delta b_m \sum_{j=1}^{s} X_{mj}^2 n_j w'_j = \sum_{j=1}^{s} X_{mj} n_j \Delta_j$$

Tables of w'_{max}, w'_{min}, Δ_{max}, and Δ_{min} are given in [6], pp. 185-188. These tables, entered with the arguments

$$Y_{j0} = b_{00} + b_{10}X_{1j} + b_{20}X_{2j} + \dots + b_{m0}X_{mj} ,$$

enable computation of $n_j w'_j$ and $n_j \Delta_j$, and therefore of the coefficients of the Δb's and the constants in (13). Equations (13) have a symmetrical matrix of coefficients and may be solved by methods used for similar simultaneous linear equation systems in multiple regressions. The process may be repeated with new provisional estimates

$$(b_{00} + \Delta b_0, b_{10} + \Delta b_1, \dots b_{m0} + \Delta b_m),$$

until the Δb's are negligible.

If the final estimates of the β's are used to evaluate the matrix of coefficients in (13), i.e., to evaluate the second derivatives of L^* at the point of maximum likelihood, the inverse of that matrix gives estimates of the variances and co-variances of the estimates of the β's: $||\sigma^2_{ik}||$, the matrix of variances and co-variances of b_i and b_k, is estimated by $||-L^*_{ik}||^{-1}$.

Testing of Hypotheses

The likelihood ratio method may be used to test hypotheses about the β's, singly and jointly. Consider, for one example, the hypothesis that the probability that $W = 1$ is independent of the values of the X's. This common probability would, in accordance with (3), be given by:

$$(14) \qquad Pr(W = 1) = Pr(\bar{I}_1 \leq \beta_0) = P(\beta_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_0 - 5} e^{-\frac{u^2}{2}} du .$$

Consequently, if the hypothesis is true, the maximum likelihood estimate of β_0 would be the value of b_0 that maximizes the following expression:

(15) $$L(b_0, 0, 0, \ldots 0) = \left(P(b_0)\right)^r \quad \left(Q(b_0)\right)^{n-r}$$

where $r = \sum_{j=1}^{s} r_j$ and $n = \sum_{j=1}^{s} n_j$.

The value b_0' which maximizes (15) is easily found to be such that:

(16) $$P(b_0') = r/n .$$

Consequently, the value of the logarithm of the likelihood function evaluated for the maximum likelihood estimate of β_0 is:

(17) $$L^*(b_0', 0, 0, \ldots 0) = r \log \frac{r}{n} + (n - r) \log \frac{n - r}{n} .$$

If the restriction of the hypothesis is removed, the maximum likelihood L^* is obtained from (6), using values of P_j and Q_j corresponding to the maximum likelihood b's. If

$$\log \lambda = L^*(b_0', 0, 0, \ldots 0) - L^*(b_0, b_1, \ldots b_m),$$

then $-2 \log \lambda$ is approximately distributed like chi-square with m degrees of freedom for large samples when the hypothesis is true.*

---

* [11], p. 259.

Other hypotheses regarding the values of β's -- for example, that $\beta_k = 0$, or that $\beta_i = \beta_k$ -- can also be tested by the likelihood-ratio method. Each test requires that the maximum likelihood estimates of the coefficients be found, and the likelihood function evaluated for these estimates, twice: both with and without the constraints implied by the hypothesis being tested. For a single hypothesis assigning definite values to all the coefficients, it may be convenient to use the hypothesized values as the initial trial values in the iterative process of finding maximum likelihood estimates. Otherwise it may be preferable to avoid the computational burden of the likelihood-ratio test by using a test based on the approximate normality of the distribution of maximum likelihood estimates from large samples: The b_k are approximately distributed by the $m + 1$ - variate joint normal distribution with means β_k and a variance-covariance matrix estimated by $||-L^*_{ik}||^{-1}$.

An Example

For purposes of illustration, an example has been worked out using data from the reinterview portion of the 1952 and 1953 Surveys of Consumer Finances conducted by the Survey Research Center of the University of Michigan for the Board of Governors of the Federal Reserve System.*

---

* A brief general description of the concepts and methods of the annual Surveys of Consumer Finances is given in [3]. For a more complete treatment, see also [10] . Reports of the 1952 and 1953 Surveys are given in [5] and [4]. I am grateful to the Survey Research Center and to the Board of Governors of the Federal Reserve System for the unpublished data used below. The paper, as well as the illustration, owes its inspiration to a semester I was enabled to spend at the Survey Research Center in 1953-1954 by the hospitality of the Center and its program of post-doctoral fellowships financed by the Carnegie Foundation.

These data were obtained from 1036 spending units who were interviewed twice, first in early 1952 and then in early 1953. The variables are as follows:

W Equal to 1 if the spending unit reported, in the 1953 interview, purchase of an automobile or any large household good (e.g., TV, washing machine, refrigerator) during 1952. Equal to 0 if the spending unit reported that it made no purchase of this kind during 1952. Spending units from whom this information was not ascertained have been omitted from the analysis.

X_1 Disposable income of the spending unit in 1952: the total income of the spending unit, as reported in the 1953 interview, less estimated income tax liability. Spending units with disposable income greater than $10,000 have been omitted from the analysis. The remainder have been classified into ten $1000-wide brackets. X_1 is taken to be the mid-point of the bracket.

X_2 Liquid asset holdings -- i.e., total of bank deposits and savings bonds -- at the beginning of 1952, as reported by the spending unit in the 1952 interview. Spending units with holdings greater than $10,000 have been omitted from the analysis. The remainder have been classified into seven categories of unequal width, as indicated in Table 1. X_2 is taken to be the midpoint of the interval.

Table 1 presents, for each pair of values (X_1,X_2), the total number of spending units included in the analysis, n_j; the number for whom $W=1$, r_j, and $W=0$, $n_j - r_j$. The frequencies in Table 1 should not be taken as representative of the population of the United States. The Surveys of Consumer Finances do collect data on distributions of income, liquid assets, and durable goods purchases that are representative of that population; tables on these distributions may be found in [4] and [5]. But the reinterview sample, on which Table 1 is based, fails to be representative insofar as it omits spending units who moved between the two surveys. Moreover, Table 1 is based on simple counts of sampled spending units, without allowance for the fact that the sampling design gave some spending units greater probabilities of being included in the sample than others. The purpose of Table 1 is not to estimate population frequency distributions, but only to examine the relationship of durable goods expenditure to income and liquid asset holdings within this sample. It is not necessary to consider here how the relationship exhibited in this sample differs from the one that would be exhibited in a complete enumeration. But it may well be that the sample gives unbiased estimates of the parameters of the relationship, even though it gives biased estimates of the separate frequency distributions of the variables.

Tables 2 - 8 give the details of the calculations. Table 2 shows the values of the coefficients b_0, b_1, and b_2 and of the corrections $\Delta b_0, \Delta b_1$, and Δb_2 in the successive iterations. The final estimate of b_1 is positive and of b_2 negative. The probability of purchasing durable goods increases with income, but decreases with liquid asset holdings. Evidently, large holders of assets are thriftier or older people, who have less inclination or need to buy durable goods. Table 3 shows for each point (X_1,X_2) the values of the index Y $(=b_0+b_1X_1+b_2X_2)$ for the initial assumed values of the b's and for the final estimates of the b's (final

Table I

Purchase of Durable Goods in Relation to Income
and Liquid Asset Holdings: 874 Spending
Units from 1952 - 1953 Surveys of Consumer Finances

Liquid Asset Holdings, Early 1952

1952 Disposable Income	X_1 \ X_2	0	1-199	200-499	500-999	1000-1999	2000-4999	5000-9999	
		0	100	350	750	1500	3500	7500	
0-999	500	1 $n_j=49$ $r_j=6$ $n_j-r_j=43$	2 6 4 2	3 5 0 5	4 7 1 6	5 10 0 10	6 7 0 7	7 5 2 3	89 13 76
1000-1999	1500	8 40 13 27	9 17 6 11	10 12 3 9	11 3 1 2	12 14 4 10	13 17 3 14	14 5 0 5	108 30 78
2000-2999	2500	15 42 15 27	16 34 13 21	17 22 9 13	18 23 11 12	19 21 9 12	20 25 7 18	21 11 2 9	178 66 112
3000-3999	3500	22 36 25 11	23 34 23 11	24 34 18 16	25 23 10 13	26 24 16 8	27 30 11 19	28 9 3 6	190 106 84
4000-4999	4500	29 23 15 8	30 22 15 7	31 21 18 3	32 26 10 16	33 12 6 6	34 39 21 18	35 5 1 4	148 86 62
5000-5999	5500	36 7 4 3	37 7 2 5	38 14 12 2	39 9 5 4	40 10 5 5	41 11 4 7	42 8 4 4	66 36 30
6000-6999	6500	43 3 2 1	44 3 0 3	45 6 4 2	46 7 7 0	47 5 1 4	48 7 2 5	49 5 3 2	36 19 17
7000-7999	7500	50 1 1 0	51 4 3 1	52 0 0 0	53 4 3 1	54 3 1 2	55 4 2 2	56 3 1 2	19 11 8
8000-8999	8500	57 0 0 0	58 2 2 0	59 2 1 1	60 3 3 0	61 6 4 2	62 6 4 2	63 2 0 2	21 14 7
9000-9999	9500	64 0 0 0	65 0 0 0	66 2 1 1	67 5 0 5	68 1 0 1	69 6 3 3	70 5 3 2	19 7 12

The number of each cell, j, is given in the upper left hand corner of the cell, the top number is n_j, the total of the other three numbers, number of spending units; the middle number is r_j, the number who made some expenditure on durable goods; the bottom number is $n_j - r_j$, the number who made no expenditure on durable goods.

iteration). Table 4 shows the matrix of coefficients and constants in the simultaneous equations (13) for the various iterations. The manner of calculation of the entries in this matrix may be illustrated, as follows. For cell 1, on iteration 1, the value of Y, Y_1, is 3.90 (see Table 3). According to Table 1, $r_1 = 6$ and $n_1 - r_1 = 43$. Entering the Table in [6], p. 185, with the value of Y_{10}, 3.90, we find: $w_1'_{min.} = .34078$ $w_1'_{max.} = .81221$ $\Delta_{1\ min.} = .25205$ $\Delta_{1\ max.} = 1.06580$. With these values, we compute $n_1 w_1' = r_1 w'_{max.} + (n_1 - r_1) w'_{min.} = 19.52680$ and $n_1 \Delta_1 = r_1 w'_{max.} - (n_1 - r_1) w'_{min.} = 1.20335$.

The matrix of coefficients in the last iteration (the first three columns of Iteration 5, Table 4) is the negative of the matrix of second derivatives of the logarithm of the likelihood function at its maximum. The inverse of this matrix gives the estimates of the variances and covariances of the b's shown in Table 5. The estimated coefficient of income X_1, b_1, is 7.4 times its estimated standard error, and the estimated coefficient of liquid asset holdings $X_2 b_2$, is 4.2 times its estimated standard error.

The hypothesis that the probability of being a buyer of durable goods is independent of both income and liquid asset holdings can be tested by the method outlined above. The total number of spending units in the sample is 874, and of these 388 were buyers while 486 were non-buyers. On the hypothesis that $\beta_1 = \beta_2 = 0$, the maximum likelihood estimate of the probab[illegible] of buying is $\frac{388}{874}$ or .444. The corresponding value of the logarithm of the likelihood function, (17), is: = -260.70858. In comparison, the logarithm of the likelihood function, (5), has at its unrestricted maximum, the value -243.622. Thus the statistic λ is $819.3 \cdot 10^{-20}$, and $-2 \log_e \lambda$ is 76.68661. By the chi-square distribution with 2 degrees of freedom, this is significant at the .95 level, and the hypothesis must be rejected.

Table 2

Iteration	b_0	Δb_1	b_1	Δb_1	b_2	Δb_2
1	3.73333	.74629	.033333	-.018877	0.00	-.0093209
2	4.47962	-.07974	.014456	.002662	-.0093209	-.0029938
3	4.39988	.009638	.017118	-.0010202	.0123147	.0038120
4	4.409518	.027442	.0160978	-.0006906	-.0085027	-.0008202
5	4.436960	-.01431	.0154072	.0006050	-.0093229	-.0002388
Final	4.42265		.0160122		-.0095617	

Table 3

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

X_1 \ X_2	0	100	350	750	1500	3500	7500
		-.0095617	.03346595	.07171275	.143415	.3346595	.7171275
500	1 3.90 4.50	2 3.90 4.49	3 3.90 4.47	4 3.90 4.43	5 3.90 4.36	6 3.90 4.17	7 3.90 3.79
1500	8 4.23 4.66	9 4.23 4.65	10 4.23 4.63	11 4.23 4.59	12 4.23 4.52	13 4.23 4.33	14 4.23 3.95
2500	15 4.57 4.82	16 4.57 4.81	17 4.57 4.79	18 4.57 4.75	19 4.57 4.68	20 4.57 4.49	21 4.57 4.11
3500	22 4.90 4.98	23 4.90 4.97	24 4.90 4.95	25 4.90 4.91	26 4.90 4.84	27 4.90 4.65	28 4.90 4.27
4500	29 5.23 5.14	30 5.23 5.13	31 5.23 5.11	32 5.23 5.07	33 5.23 5.00	34 5.23 4.81	35 5.23 4.43
5500	36 5.57 5.30	37 5.57 5.29	38 5.57 5.27	39 5.57 5.23	40 5.57 5.16	41 5.57 4.97	42 5.57 4.59
6500	43 5.90 5.46	44 5.90 5.45	45 5.90 5.43	46 5.90 5.39	47 5.90 5.32	48 5.90 5.13	49 5.90 4.75
7500	50 6.23 5.62	51 6.23 5.61	52 6.23 5.59	53 6.23 5.55	54 6.23 5.48	55 6.23 5.29	56 6.23 4.91
8500	57 6.57 5.78	58 6.57 5.77	59 6.57 5.75	60 6.57 5.71	61 6.57 5.64	62 6.57 5.45	63 6.57 5.07
9500	64 6.90 5.94	65 6.90 5.93	66 6.90 5.91	67 6.90 5.87	68 6.90 5.80	69 6.90 5.61	70 6.90 5.23

In each cell the upper number is the value of Y for the initially assumed values of b_0, b_1, and b_2 (iteration 1, Table 2), and the lower number is the value of Y for the final estimates of b_0, b_1, and b_2 (last row, Table 2).

Table 4

$\sum_{j=1}^{s} n_j w'_j$			$\sum_{j=1}^{s} n_j \Delta_j$
$\sum_{j=1}^{s} X_{1j} n_j w'_j$	$\sum_{j=1}^{s} X^2_{1j} n_j w_j'$		$\sum_{j=1}^{s} X_{1j} n_j \Delta_j$
$\sum_{j=1}^{s} X_{2j} n_j w'_j$	$\sum_{j=1}^{s} X_{2j} X_{1j} n_j w_j'$	$\sum_{j=1}^{s} X^2_{2j} n_j w_j'$	$\sum_{j=1}^{s} X_{2j} n_j \Delta_j$

Iteration 1

	b_1	b_2	=
506.26226			- 36.18095
18,297.96880	850,885.5965		-5273.56185
7358.683175	307,558.8742	316,137.4833	-3261.31942

Iteration 2

	b_1	b_2	=
531.68769			- 8.64421
19,078.17335	903,006.3782		155.60315
7409.98837	316,237.7764	310,961.1571	-463.98057

Iteration 3

525.44381			13.23685
18,917.35435	895,898.72925		435.18985
7206.56012	306,093.18985	298,937.514995	896.75023

Iteration 4

529.88552			- 4.69292
19,044.34130	900,917.4600		-366.59100
7415.665625	325,587.2889	312,111.16541	-277.34406

Iteration 5

533.15147			2.29643
19,335.50135	928,779.8088		208.89305
7431.27326	320,036.8671	310,559.7028	13.16229

Table 5

Estimates of Variances and Covariances of Coefficients.
(Negative of Inverse of Matrix of Second Derivatives of Logarithm of Likelihood Function, Evaluated at Point of Maximum Likelihood.)

+ .007832610990		
- .0001527020365	+ .00000464654363	
- .000030061182432	- .000001134386407	+ .000005108335533

The high value of b_2 relative to its estimated standard error indicates that the hypothesis that $\beta_2=0$ can be rejected. This hypothesis can also be tested by the likelihood ratio method. Table 6 reports the results of a series of iterations to obtain maximum likelihood estimates of b_0 and b_1, on the assumption that $b_2=0$. Table 7 shows the values of Y for the first and last of these iterations. The third column of Table 7 shows the observed values of $n_j, \frac{r_j}{n_j}$, and $\frac{n_j - r_j}{n_j}$ for each of the ten levels of income. The fourth column shows, for each level of income, the "predicted" probability of buying P_j and of not buying, Q_j. P_j is the value of the cumulative unit-normal distribution function corresponding to the final iteration Y_j shown in the second column. Q_j is $1 - P_j$. From these values, the

Table 6

Iteration	β_0	$\Delta\beta_0$	β_1	$\Delta\beta_1$
1a	3.73333	.68260	.033333	-.020918
2a	4.41593	-.05506	.012415	+.001537
3a	4.36087	.00137	.013952	-.000066
4a	4.36224	-.00159	.013886	+.000047
5a	4.36065	-.00036	.013933	-.000005
6a	4.36029	-.00089	.013928	.000015
7a	4.35940	-.00036	.013943	-.000005
8a	4.35904	-.00089	.013938	.000015
Final	4.35815		.013953	

logarithm of the likelihood function, (5), can be evaluated at its maximum for $\beta_2=0$. Comparing this with the unrestrained maximum, gives a likelihood ratio λ of $253.4\ 10^{-10}$. $-2 \log \lambda$ is equal to 34.99, which is significant at .95 level according to the table of chi-square for 1 degree of freedom.

The example has been presented for illustration of a method rather than for substantive results. Still more variables would be needed for a better explanation of durable goods purchasing behavior. Moreover, it is wasteful of information to disregard amounts spent by those who purchased. Some combination of probit analysis and regression is indicated, to handle a variable with a large probability of having zero value and the remaining probability spread over a positive interval.

Table 7

Income X_1		$Y=b_0+b_1X_1$ Iteration 1	Final Iteration		n_j	$\frac{r_j}{n_j}$	$\frac{n_j - r_j}{n_j}$	P_j	Q_j
500	1	3.90	4.43	1	89	.1461	.8539	.2843	.7157
1500	2	4.23	4.57	2	108	.2778	.7222	.3336	.6664
2500	3	4.57	4.71	3	178	.3708	.6292	.3859	.6141
3500	4	4.90	4.85	4	190	.5579	.4421	.4404	.5596
4500	5	5.23	4.99	5	148	.5811	.4189	.4960	.5040
5500	6	5.57	5.13	6	66	.5455	.4545	.5517	.4483
6500	7	5.90	5.27	7	36	.5278	.4722	.6064	.3936
7500	8	6.23	5.40	8	19	.5789	.4211	.6554	.3446
8500	9	6.57	5.54	9	21	.6667	.3333	.7054	.2946
9500	10	6.90	5.68	10	19	.3684	.6316	.7517	.2483

References

[1] Aitchison, J. and Brown, J. A. C., "An Estimation Problem in Quantitative Assay," Biometrika, 41 (1954), 338 - 343.

[2] Aitchison, J. and Brown, J. A. C., "A Synthesis of Engel Curve Theory," Review of Economic Studies, 22 (1955), 35 - 46.

[3] Board of Governors of the Federal Reserve System, "Methods of the Survey of Consumer Finances," Federal Reserve Bulletin, July 1950.

[4] Board of Governors of the Federal Reserve System, 1953 Survey of Consumer Finances, reprinted with supplementary tables from Federal Reserve Bulletin, March, June, July, August, and September 1953.

[5] Board of Governors of the Federal Reserve System, 1952 Survey of Consumer Finances, reprinted with supplementary tables from Federal Reserve Bulletin, April, July, August, and September 1952.

[6] Cornfield, J. and Mantel, N., "Some New Aspects of the Application of Maximum Likelihood to the Calculation of the Dosage Response Curve," Journal of the American Statistical Association, 45 (1950), 181 - 210.

[7] Farrell, M. J., "The Demand for Motor Cars in the United States," Journal of the Royal Statistical Society, Series A, 117, Part 2 (1954) 171 - 193.

[8] Finney, D. J., Probit Analysis, 2nd. edition, Cambridge, England: the University Press, 1952.

[9] Garwood, F. "The Application of Maximum Likelihood to Dosage Mortality Curves," Biometrika, 32 (1941), 46 - 58.

[10] Klein, L. R., editor, Contributions of Survey Methods to Economics, New York: Columbia University Press, 1954.

[11] Mood, A. M., Introduction to the Theory of Statistics, New York: McGraw-Hill, 1950.