Contributions to Methodology and Statistics
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 10, Ljubljana: FDV 1995

Comparison of the Logistic Regression Model and the Linear Probability Model of Categorical Data

Cveto Trampuž[1]

Abstract

The aim of this paper is to examine in what situations the interpretation of results involving empirical data may differ, if in the analysis of mutual relationships between all nominal variables, either a linear probability model or a logistic regression model is used. An example is given where such differences are possible if some of the conditional ratios of frequencies ($f_{y=1}/f_{y=0}$) of the dependent variable y, at given values of the independent variables, differ greatly from 1.

Keywords: Logistic Regression; Linear probability model; Categorical Data.

1 Introduction

1.1 Definition of the problem

In the investigation of the relationship between nominal variables various statistical models can be used in the examination of the same problem using the same data. The question which arises is whether the inferences regarding the relationship between the variables considered may differ depending on the model used.

In this paper two models are compared: the linear probability model (with the ordinary last squers method of estimating) and the logistic regression model (with the maximum likelihood method of estimating) in which the independent variables may have many different values. The findings for the logistics regression model directly apply to the equivalent logit model.

It is known that selecting either of the two models may lead to different conclusions regarding the relationships between nominal variables, if some of the

[1] Faculty of Social Sciences, University of Ljubljana, P.O Box 47, 61109 Ljubljana, Slovenia

conditional probabilities are greater than 0.75 or less than 0.25 (Goodman, 1978). The question of how precise the boundaries of the mentioned interval are, remains open.

The primary purpose of this paper is to find an appropriate method of comparison which allows the most general conclusions possible to be drawn.

One method of comparison is to find explicit formulas for parameter estimates for both models as functions of the same empirical data and to examine how the estimates of the parameters of the models formally vary, depending on the formula used and depending on the variation in all or only some of the data.

This paper is not concerned with the appropriateness of the model in relation to the meaning of the variables. Similarly, it does not deal with the problems of statistical assessment of computed parameter estimates.

The simplest case, with one dependent and two independent variables (of which one has two and the other three values), is used as the illustrative example. The findings may of course be generalised to models with any number of independent variables.

1.2 Presentation of given data in two-dimensional cross-classification table

Multidimensional frequency tables may be expressed in the form of a two-dimensional table where rows represent the values of the dependent variable and columns represent a combination of the values of the other (independent) variables. The frequencies in such a table are the initial input data for all further computations.

As an illustrative example, such a table is presented in the case of three variables.

Suppose we have a sample of N independent observations (y_i, s_i, z_i), where y_i denotes the value (coded as 0 or 1) of the dichotomous dependent variable Y, s_i is the value of the independent variable S (coded as 0 or 1) and z_i is the value of the independent variable Z (coded as 1 or 2 or 3) for the i^{th} subject ($i = 1, 2, \ldots, N$). The dummy dichotomous variable is determined for each value of Z by the usual procedure and denoted sequentially as Z1 and Z2:

$$Zk = \begin{cases} 1 & \text{if } Z = k \\ 0 & \text{if } Z \neq k \end{cases}$$

The data under consideration are the frequencies given in a three-dimensional table which is presented in the form of a two-dimensional Table 1.

The second row reading across columns shows how all the possible combinations of the values of variables S and Z form the cells of the table (in any

sequence desired, and denoted by just one index j) and hence the meaning of the frequency f_{kj}.

The table is introduced primarily in order to determine the quantities and their denotations for use in the subsequent computations. The meaning of the notations in Table 1 are:

f_{kj} is the frequency of cell kj, (k= 0,1; j = 1,2,...,6),

$$f_{k+} = \sum_{j} f_{kj} ,$$

$$f_{+j} = \sum_{k} f_{kj} ,$$

$$N = \sum_{k}\sum_{j} f_{kj} = \sum_{k} f_{k+} = \sum_{j} f_{+j} .$$

Table 1. Presentation of a three-dimensional table in two dimensions

			S	Z			
	01	02	03	11	12	13	
j	1	2	3	4	5	6	
Y=0	f_{01}	f_{02}	f_{03}	f_{04}	f_{05}	f_{06}	f_{0+}
Y=1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{1+}
	f_{+1}	f_{+2}	f_{+3}	f_{+4}	f_{+5}	f_{+6}	N

2 Linear probability model

The general linear probability model is expressed as:

$$y_i = \sum_{j=1}^{J} b_j x_{ij} + u_i , \quad (i=1,2,...,N) \tag{1}$$

where

- Y is the dependent variable (y_i equals either zero or one);
- X_j denotes the independent variables, j = 2,3,...,J (which themselves have two values or are dummy variables derived from nominal polytomous variables) which may also be all possible products of the variables. The values of the independent variables (including X_1) for i^{th} subject shall be denoted as x_i ($x_{i1} \equiv 1$);
- b_j unknown and sought parameters of the model;
- u error term;
- i index denoting the i^{th} subject from the sample of N size;

J number of all variables in the model.

The expected value of variable y_i (denoted as P_i) is the probability that y_i equals one for the given values of the independent variables.

$$P_i = E(y_i) = P(y_i = 1/x_i) = \sum_{j=1}^{J} b_j x_{ij}$$

The equations for the computation of the estimates of the parameters are:

$$\sum_{i=1}^{N}(y_i - p_i) = 0$$
$$\sum_{i=1}^{N} x_{ij}(y_i - p_i) = 0 \ , \ j=2,3,\ldots,J. \qquad (2)$$

It is evident from Table 1 that some subjects have the same value of x_i. Let J denote the number of distinct values among x_i. The number of subjects with $x_i=x_j$ will be denoted by F_{+j} ($j=1,2,\ldots,J$; $\Sigma F_{+j} = N$). Let F_{1j} denote the number of 'responses', $y=1$, among the F_{+j} subjects with $x_i=x_j$ ($\Sigma F_{1j} = F_{1+}$). F_{1+} is the total number of subjects with $y=1$. It follows that there are J distinct values of P_i only.

The estimate for distinct values of probability P_j is denoted by p_j and the estimate for b_j is denoted by B_j ($j=1,2,\ldots,J$).

For the purpose of further comparison with the logistic regression model the equations (2) are modified into a form which makes it explicitly evident how the parameters B_j and p_j are dependent on the empirical data F_{+j} and F_{1j} both for the saturated model as well as any other feasible non-saturated model.

The equations (2) may be expressed in a matrix form (derivation is simple, but because it takes up a lot of space it is not presented here). All matrices are of order JxJ:

$$E_1D^TF_1 = E_1D^TFp$$
$$O = E_2D^{-1}p, \qquad (3)$$

where

D is the 'design' matrix for the saturated model (where the values of the independent variables are coded as 1 and 0). It is determined such that the equation:

$$p = DB$$

is valid;

E_1 is the diagonal matrix which has zeroes or ones along the diagonal. At those points where the saturated model parameters b_j are set at 0 (in order

to define a nonsaturated model) the diagonal elements of matrix E_1 are set at 0. By means of E_1 then each feasible non-saturated model is defined. In the case of a saturated model all E_1's diagonal elements are equal to 1;

E_2 is the diagonal matrix whose diagonal elements are zeroes except at the points where the parameter b_j is set at 0. In the case of a saturated model E_2 has only zeroes on the diagonal ($E_1 + E_2$ = identity matrix);

p is the vector of p_j: $p^T = (p_1, p_2, \ldots, p_j)$;

e is the vector of ones only: $e^T = (1, 1, \ldots, 1)$;

O is the vector of zeroes only: $O^T = (0, 0, \ldots, 0)$;

F_1 is the vector with values f_{1j} which have the same meaning as in Table 1:
$F_1^T = (f_{11}, f_{12}, \ldots, f_{1j})$;

F is the diagonal matrix with elements f_{+j} which have the same meaning as in Table 1:

$$F = \begin{matrix} f_{+j} & 0 & 0 & \ldots & 0 \\ 0 & f_{+2} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & f_{+J} \end{matrix}$$

Equations (3) may be combined or 'added' (the second part of the equation is inserted in the first part; matrices E_1 and E_2 assure correct 'addition') and may be expressed as a single system of linear equations:

$$E_1 D^T F_1 = (E_1 D^T F + E_2 D^{-1}) p \qquad (4)$$

Let as assume that the independent variables are not linearly interdependent. Without providing a proof it may be noted that the matrix $(E_1 D^T F + E_2 D^{-1})$ has a rank equal to the number of the unknown, hence p may be determined from the equation (4).

Matrix H is introduced to shorten the presentation of the solution of the equation (4):

$$H = (E_1 D^T F + E_2 D^{-1})^{-1} E_1 D^T .$$

The equation for p then is:

$$p = H F_1 .$$

Since the equation:

$$DB = p ,$$

also holds, the computation of B is expressed as:

$$B = D^{-1}p \quad \text{or} \quad B = D^{-1}HF_1 . \tag{5}$$

Using the same model (defined by E_1 and E_2) and with constant frequencies F, matrices H and D^{-1} H show how the values of p or B vary with changes in frequencies F_1. The F_1 frequencies can only be changed at the $0 < F_1 < F$ interval. The frequencies F may also be changed, but then the matrix H must be recomputed.

NUMERICAL EXAMPLE

Table 2: Example of hypothetical data

	SZ						
	01	02	03	11	12	13	
j =	1	2	3	4	5	6	
Y= 0	37	11	51	62	19	65	245
Y= 1	136	61	170	173	80	176	796
	173	72	221	235	99	241	1041

An example with only three nominal variables is given. Any conclusions may of course be generalised to a model with more nominal variables.
The saturated model has the following form:

$$Y = b_1 + b_2S + b_3Z1 + b_4Z2 + b_5SZ1 + b_6SZ2 + u, \tag{6}$$

where the symbols have the same denotation as in the general model (1) described. Index i has been dropped to simplify the notation.

The computation of p_j and B_j is presented in matrix form for saturated and any feasible nonsaturated model from (6).

The values of the matrices and vectors are presented for the saturated model (6) first. These matrices and vectors are constant for all non-saturated models. The values of F_1 and F are taken from Table 2.

$$D = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad D^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{bmatrix}$$

The method of computing the design matrix D is simple. As it is not pertinent here it is described only briefly:
We proceed from the model (6). For every combination of values of the independent variables S and Z (there are 6 such combinations in the present example) first the values of the variables S, Z1 and Z2 are determined (columns 2,3 and 4 of D) and then by multiplying S by Z1 and S by Z2 the variables SZ1 and SZ2 are constructed (column 5 and 6), so that the equation $p=DB$ is valid. In the first column of matrix D there are only ones.

The matrix D^{-1} is only presented here as an example of the validity of the assertion that the sum of the elements by rows equals 0 (except for the first row which determines the constant parameter of the model).

$$F = \begin{matrix} 173 & 0 & 0 & 0 & 0 & 0 \\ 0 & 72 & 0 & 0 & 0 & 0 \\ 0 & 0 & 221 & 0 & 0 & 0 \\ 0 & 0 & 0 & 235 & 0 & 0 \\ 0 & 0 & 0 & 0 & 99 & 0 \\ 0 & 0 & 0 & 0 & 0 & 241 \end{matrix}$$

$F_1^T = (136, 61, 170, 173, 80, 176)$,
$e^T = (1,1,1,1,1,1)$,
$p^T = (p_1,p_2,p_3,p_4,p_5,p_6)$ is the vector of the unknown estimates of P_j,
$B^T = (B_1,B_2,B_3,B_4,B_5,B_6)$ is the vector of the unknown estimates of b_j.

Equation (3) holds for the saturated model as well as for all feasible non-saturated models with the selected variables in model (6). Non-saturated models are defined only by selection of different combinations of ones and zeroes on the diagonal matrix E_1 and E_2.

For example consider the following non-saturated model:

$$Y = b_1 + b_2S + b_3Z1 + b_4Z2 + u\,, \qquad (7)$$

The values of matrix E_1 and E_2 on the above model are:

$E_1 = \mathrm{diag}(1, 1, 1, 1, 0, 0)$
$E_2 = \mathrm{diag}(0, 0, 0, 0, 1, 1)$.

Matrix H multiplied by 100:

$$100H = \begin{matrix} 0.3744 & 0.1299 & 0.1171 & 0.1499 & -0.0945 & -0.1074 \\ 0.1299 & 0.7154 & 0.1177 & -0.0957 & 0.4898 & -0.1079 \\ 0.1171 & 0.1171 & 0.3225 & -0.0862 & -0.0856 & 0.1192 \\ 0.1499 & -0.0957 & -0.0856 & 0.3152 & 0.0696 & 0.0790 \\ -0.0945 & 0.4898 & -0.0856 & 0.0696 & 0.6539 & 0.0785 \\ -0.1074 & -0.1079 & 0.1192 & 0.0790 & 0.0785 & 0.3056 \end{matrix}$$

Matrix D^{-1} H multiplied by 100:

$$100D^{-1} = \begin{matrix} -0.1074 & -0.1074 & 0.1192 & 0.0790 & 0.0785 & 0.3056 \\ 0.2245 & 0.2256 & 0.2033 & -0.1652 & -0.1641 & -0.1864 \\ 0.2573 & 0.0123 & -0.2054 & 0.2361 & -0.0089 & -0.2266 \\ 0.0129 & 0.5977 & -0.2048 & -0.0095 & 0.5754 & -0.2271 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{matrix}$$

Parameters p and B:

j	p	b
1	0.7823	0.7282
2	0.8496	0.0433
3	0.7715	0.0108
4	0.7390	0.0781
5	0.8064	0.0000
6	0.7282	0.0000

3 Logistic regression model

Let us assume that for the sample size of N observations we have the values for the dependent variable Y (with values 0 and 1) and m-1 independent nominal polytomous variables X_2, X_3 .... , X_m. In the computations their internal values (the values of design variables) are determined by the method of coding. Hereafter the same coding scheme is considered as in the linear probability model.

To present the model and formula for computing the estimates of the model parameter the following symbols and quantities are used (to simplify the notation, index i which determines the index of the subjects, is omitted wherever it is not essential):

$\beta^T = (\beta_1, \beta_2, \ldots, \beta_m, \ldots, \beta_J)$, vector of unknown parameters β_j,

$X^T = (X_1, X_2, \ldots, X_m, \ldots, X_J)$, vector of independent variables, with $X_{m+1},\ldots,X_J$ denoting the possible mutual products of the independent variables

$X_2, X_3,\ldots,X_m$. All values of 'variable' X_1 are ones. The values of all the independent variables for the i^{th} subject shall be denoted as x_i ($x_{i1} \equiv 1$).

The unknown conditional probability of the event $P_i(Y=1/x_i)$ is denoted by π_i:

$$\pi_i = E(y_i/x_i) = P(y_i=1/x_i) ,$$

the estimate of π_i is denoted by p_i'.

Further, let us introduce

$$\gamma_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_J X_{iJ} \text{ and}$$

$$\pi_i = \exp(\gamma_i)/[1+\exp(\gamma_i)] .$$

The usual equation for computing the parameters of the logistic regression model under the maximum likelihood method is:

$$\begin{aligned} &\sum_{i=1}^{N}(y_i - p_i) = 0 \\ &\sum_{i=1}^{N} x_{ij}(y_i - p_i) = 0 , \quad j=2,3,\ldots,J. \end{aligned} \tag{8}$$

It has been already noted that some subjects have the same value of x_i. Let J denote the number of distinct values among x_i . The number of subjects with $x_i=x_j$ (j=1,2,...,J) will be denoted by F_{+j} ; $\Sigma F_{+j} = N$). Let F_{1j} denote the number of 'responses', y=1, among the F_{+j} subjects with $x_i=x_j$ ($\Sigma F_{1j} = F_{1+}$ is the total number of subjects with y=1). It follows that there are J distinct values of π_i or γ_i only. Let p_j' (j=1,2,...,J) denote the estimate of the distinct values of probability π_i , B_j' denote the estimate of parameters β_j and g_j denote the estimate of the distinct values of γ_i.

The equations (8) may be expressed in a matrix form that is the same as that introduced into the linear probability model (3):

$$\begin{aligned} E_1 D^T F_1 &= E_1 D^T F p' \\ O &= E_2 D^{-1} g . \end{aligned} \tag{9}$$

Matrices D, E_1, E_2, F and vectors F_1 and O have exactly the same meaning as in the linear probability model (3).

p' is the vector of p_j': $p'^T = (p_1', p_2', \ldots, p_J')$;
g is the vector of g_j: $g^T = (g_1, g_2, \ldots, g_J)$;
B' is the vector of B_j': $B'^T = (B_1', B_2', \ldots, B_J')$.

The meaning of p' is identical to that of p in (3). Consequently p and p' are directly comparable quantities.

In the equations (9) a feasible non-saturated logistic regression model may be defined with an appropriate selection of zeroes and ones in matrices E_1 and E_2.

Matrix D is a design matrix of the saturated logistic regression model. It may be determined in the same way as that described before. It is almost always given in computations in the framework of computer program packages.

For logistic regression models the equation holds:

$$g = DB'.$$

Once g has been computed, B' may be obtained by:

$$B' = D^{-1}g. \tag{10}$$

It may be observed that equations (9) are similar to equations (3) except that on the right-hand side of the second part of equations (9) it is not directly p' but g which is a non-linear function of p':

$$g_j = \log[p_j'/(1-p_j')].$$

From the equations (5) and (10) can be seen, that parameters B are the function of the differences $P_j - P_{j'}$ ($j \neq j'$) and parameters B' are the functions of the differences $\gamma_j - \gamma_{j'}$ at the same indexes j and j', determined by the matrix D^{-1}.

The first part of the equations (9) are linear equations for computing p' just as the first part of equations (3) are for computing p.

The second part of the equations (9) are polynomials of p_j' which may decompose into linear factors depending on the model.

The logistic regression model includes the log ratio: $g_j/2$. It is shown in Table 3 that the difference: $[g_j/2 - (2p_j'-1)]$ is small at $p_j' \sim (1/2)$. Consequently the expression $g_j/2$ may be substituted by linear approximation $2p_j'-1$ ($j=1,2,\ldots,J$) in the second part of the equations (9).

To demonstrate this the Taylor exponential series is presented:

$$(1/2)\log[r/(1-r)] \sim (2r-1) + (2r-1)^3/3 + (2r-1)^5/5 + \ldots.$$

Consider the linear term only and find the extent to which these two quantities differ for various values of r:

Table 3. The difference between (1/2)log[r/(1-r)] and (2r-1) for various values of r.

r	(1/2)log[r/(1-r)]	(2r-1)	Absolute error
0.50	0.00000	0.00000	0.00000
0.60	0.20273	0.20000	0.00273
0.65	0.30952	0.30000	0.00952
0.70	0.42365	0.40000	0.02365
0.75	0.54931	0.50000	0.04931
0.80	0.69315	0.60000	0.09315
0.85	0.86730	0.70000	0.16730
0.90	1.09861	0.80000	0.29861
0.99	2.29756	0.98000	1.31756

It is evident from Table 3 that the value of (1/2)log[r/(1-r)] and (2r-1) begin to differ to a greater degree when the value of r is in the interval $0.75 < r < 0.25$. How the error of these approximations influences errors in solving the equations (11) remains an open question. For more precise conclusions this problem should be examined by numerical analysis methods.

To compare further the results of the two models under examination, in the second part of the equations (9) vector g is replaced by the linear approximation 2(2p'-e).

Taking into account the fact that

$$E_2D^{-1}\,C = 0 \text{ holds,}$$

where C is the vector with all elements equal to any constant c, we get a linear system of equations for computation of the approximation of p'. This approximation is denoted by p":

$$E_1D^TF_1 = E_1D^TFp"$$
$$O = E_2D^{-1}p" . \qquad (11)$$

p" may be computed from the above equations by the same procedure used with the linear probability model. The first and second part of equation (11) may be added and matrix H defined:

$$H = (E_1D^TF + E_2D^{-1})^{-1}\,E_1D^T.$$

For computation of p" the following formula is valid:

$$p" = HF_1 .$$

Parameter B" is computed by the formula:

$$B'' = D^{-1}g'' ,$$

where

$$g_j'' = \log[p_j''/(1-p_j'')].$$

Equations (9) allow certain conclusions concerning the comparison of the results obtained by the two methods.

For equations (9) it is also quite simple to make an iteration procedure for a computation of p' and B'.

NUMERICAL EXAMPLE

If we take the matrices and vectors for the logistic model similar to those of the probability linear model and the same data from Table 2, then

p" and B" are computed by approximation (11),

p' and B' are computed by the iteration procedure from equations (9):

Table 4: Parameters p', p", B' and B":

j	p'	p"	B'	B"
1	0.7828	0.7823	0.9801	0.9857
2	0.8446	0.8496	0.2433	0.2310
3	0.7727	0.7715	0.0586	0.0552
4	0.7386	0.7390	0.4696	0.4408
5	0.8100	0.8064	0.0000	0.0070
6	0.7271	0.7282	0.0000	0.0740

4 Conclusions

4.1 The form of the equations (3) and (9) makes it possible to determine with both the logistic and the probability linear regression models (for a saturated and all feasible non-saturated models) how the values of computed parameters vary if the values of F_1 and F are changed.

4.2 The second part of the equations (9) includes a log ratio: $g_j/2$. The quantities $g_j/2$ and $(2p_j'-1)$ are approximately the same for values of p_j' at the interval approximately from 0.25 to 0.75 . Consequently the expression $g_j/2$ may be substituted by the expression $2p_j'-1$ in the second part of the equations (9). If the computed values for p_j' are at the interval cited, then the results obtained for the

two models are approximately the same and do not allow different inferences regarding the relationship between the examined variables to be made. Practical examples show, however, that the results obtained by applying either of the two models do not differ, although some of estimates p_j' are outside the mentioned range (Goodman, 1976). A detailed numerical analysis of equations (9) would provide a more exact answer as to whether the rang 0.25 to 0.75 can be changed. If the value of p_j' differs little for the different j so that $p_j' = c \pm \varepsilon_j$, $[\varepsilon_j < c(1-c)]$ is valid, where c is any constant at the interval: $0 < c \pm \varepsilon < 1$ then constant c in the second part of equations (9) influences the result only minimally. Only the differences between ε_j are important. The errors are minimal at $c \sim 1/2$.

4.3 The second part of the equations (9) are polynomials of p_j' which may decompose into linear factors depending on the model. In the case that second part of the equations (9) break up into linear factors (in such cases there exist explicit expression for estimations p'), the equations (9) and the equations (3) are equivalent. It is also known that the estimations of p and p' are always the same for the saturated model.

4.4 Differences in the interpretation of the results obtained by use of either model on the same data may therefore be expected to arise when the odds ratios: $[p_j'/(1-p_j')]/[p_{j'}'/(1-p_{j'}')]$, $j',j=1,2,\ldots,J$ differ considerably from 1 at least for a pair j' and j ($j \neq j'$) which has a great (numerically critical) influence on solving equations (9). In a similar situation the results obtained by the linear probability model might be incorrect. Estimates for P might be greater than 1 or negative. Practical examples show, that the logistic regression models fit the data better when some odds ratios differ considerably from 1.

5 Numerical example

We have selected the following model as an example:

$$Y = b_1+b_2S+b_3Z1+b_4Z2+b_5T1+b_6T2+b_7T3+b_8SZ1+b_9SZ2+b_{10}ST1+b_{11}ST2 +b_{12}ST3+b_{13}Z1T1+b_{14}Z1T2+b_{15}Z1T3+b_{16}Z2T1+b_{17}Z2T2+b_{18}Z2T3 + u ,$$

where in addition to variables Y, S and Z, used for the accompanying example, variable T which has four values (1,2,3,4) was introduced. Similarly as in variable Z, dichotomous variables T1, T2 and T3 with the values 0 and 1 were derived from variable T. Variables SZ1, ..., Z2T3 are the products of dichotomous variables S, Z1 and T3 , generated in the same way as described in Section 1 for model (6). We do not present the design matrix because of its size.

The data is fictitious and selected in such a way that some ratios f_{0j}/f_{1j} differ greatly from 1, so that possible differences in the interpretation of results obtained with one or the other model could be demonstrated.

Table 5: Computed parameters p and B for the linear probability model, and p' and B' for the logistic regression model

SZT	f_{+j}	f_{1j}	f_{1j}/f_{+j}	f_{0j}/f_{1j}	p	$f_{1j}/(pf_{+j})$	p'	$f_{1j}/(p'f_{+j})$
011	57	12	0.211	3.75	0.1376	1.53	0.1411	1.49
012	59	12	0.203	3.92	0.1620	1.26	0.1692	1.20
013	24	7	0.292	2.43	0.2131	1.37	0.2072	1.41
014	33	1	0.030	32.00	0.2875	0.11	0.2728	0.11
021	21	2	0.095	9.50	0.0553	1.72	0.0903	1.05
022	24	6	0.250	3.00	0.2288	1.09	0.2315	1.08
023	12	2	0.167	5.00	0.1152	1.45	0.1345	1.24
024	15	1	0.067	14.00	0.1977	0.34	0.1289	0.52
031	68	13	0.191	4.23	0.2646	0.72	0.2509	0.76
032	67	16	0.239	3.19	0.2829	0.84	0.2756	0.87
033	33	6	0.182	4.50	0.2576	0.71	0.2549	0.71
034	53	52	0.981	0.02	0.7839	1.25	0.8125	1.21
111	83	23	0.277	2.61	0.3272	0.85	0.3248	0.85
112	66	14	0.212	3.71	0.2492	0.85	0.2427	0.87
113	31	10	0.323	2.10	0.3834	0.84	0.3880	0.83
114	55	15	0.273	2.67	0.1184	2.30	0.1272	2.14
121	36	7	0.194	4.14	0.2177	0.89	0.1973	0.99
122	26	7	0.269	2.71	0.2888	0.93	0.2863	0.94
123	14	3	0.214	3.67	0.2584	0.83	0.2418	0.89
124	23	2	0.087	10.50	0.0015	57.32	0.0464	1.87
131	98	31	0.316	2.16	0.2654	1.19	0.2749	1.15
132	72	16	0.222	3.50	0.1812	1.23	0.1880	1.18
133	23	8	0.348	1.88	0.2390	1.46	0.2429	1.43
134	48	10	0.208	3.80	0.4261	0.49	0.3945	0.53

The indication SZT in the first column of Table 5 above denotes combinations of values of variables S, Z, and T and hence the meaning of frequencies f_{1j} and f_{+j} in the two-dimensional table made out of the four-dimensional table, similar as described for Table 1.

In Table 6 symbol sig(T) denotes significance of statistics T for the linear probability model, symbol sig(W) denotes significance of Wald's statistics for the logistic regression model. On the basis of these two statistics we might make some inferences regarding the relationship between the examined variables.

The example presented indicates that the interpretation of results based on statistics sig(T) and sig(W) might differ in variables T1, T3 and XZ2.

Estimations for p and p' do not differ considerably. Differences occur mostly where p is small (table cells: 021 and 124).

Table 6

Variable	B	sig(T)	B'	sig(W)
C	0.4261	0.000	-0.4285	0.116
X	0.3579	0.001	1.8951	0.000
Z1	-0.3076	0.000	-1.4971	0.000
Z2	-0.4245	0.000	-2.5946	0.000
T1	-0.1607	0.012	-0.5414	0.108
T2	-0.2449	0.000	-1.0345	0.005
T3	-0.1870	0.032	-0.7081	0.139
XZ1	-0.1889	0.001	-0.9501	0.005
XZ2	-0.1617	0.034	-0.7828	0.113
XT1	-0.3586	0.000	-2.0192	0.000
XT2	-0.2562	0.001	-1.3987	0.002
XT3	-0.3393	0.000	-1.8310	0.001
Z1T1	0.3695	0.000	1.7351	0.000
Z1T2	0.3756	0.000	1.8222	0.000
Z1T3	0.4520	0.000	2.1779	0.000
Z2T1	0.3769	0.000	2.1614	0.006
Z2T2	0.5322	0.000	3.1442	0.000
Z2T3	0.4439	0.001	2.5885	0.004

References

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

[2] Goodman Leo A. (1978): *Analyzing Qualitative / Categorical Data,* Cambridge: Abt Books.

[3] Haberman Shelby J. (1978 and 1979). *Analysis of Qualitative Data. Volume 1, 2,* New York: Academic press.

[4] Dobson Annette J. (1983): *Introduction to Statistical Modelling.* London: Champman and Hall.

[5] Agresti Alan (1990): *Categorical Data Analysis.* New York: Wiley.

[6] Hosmer David W. and Stanley Lemeshow (1980): *Applied Logistic Regression.* New York: Wiley.