MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 15, NUMBER 1, SUMMER 1986

Using Multiple Regression with Dichotomous Dependent Variables

Jerome D. Thayer
Andrews University

Introduction

Dichotomous variables are frequently encountered in multiple regression analysis, both as independent and dependent variables. A dichotomous independent variable is used to determine whether group membership is related to or will predict a certain outcome (i.e., whether gender predicts gpa). A dichotomous dependent variable is used to determine a combination of variables that will predict group membership (i.e., to predict dropping out of college).

Historically, whenever a dichotomous variable was studied as an independent variable with one dependent variable, a t-test, analysis of variance or analysis of covariance was conducted. When a dichotomous variable was studied as a dependent variable, discriminant analysis was used.

As multiple regression became more common, its advocates suggested that it could or should replace the t-test, ANOVA, ANCOVA or discriminant analysis in dealing with dichotomous variables by using coded variables.

Recently, however, Cox (1970), Goodman, (1978), Aldrich and Nelson (1984), and others have questioned the practice of using multiple regression when a dichotomous variable is used as the dependent variable. The most frequently suggested replacement for multiple regression is logistic regression.

In the introduction to Aldrich and Nelson (1984), it is suggested that ordinary regression analysis is not an appropriate strategy to analyze qualitative dependent variables, including those that are dichotomous. They go on to express the limitations of multiple regression very strongly:

> Perhaps because of its widespread popularity, regression may be one of the most abused statistical techniques in the social sciences. While estimates derived from regression analysis may be robust against errors in some assumptions, other assumptions are crucial, and their failure will lead to quite unreasonable estimates. Such is the case when the dependent variable is a qualitative measure rather than a continuous, interval measure. . . . For example we shall show that regression estimates with a qualitative dependent variable may seriously misestimate the magnitude of the effects of independent variables, [and] that all of the standard statistical inferences such as hypothesis tests . . . are unjustified (p. 9, 10).

The authors suggest that the failure of regression is "particularly troubling in the behavioral sciences" (p. 10), giving examples of qualitative dichotomous variables from the fields of political science, economics and sociology. Similar criticisms concerning dichotomous dependent variables are given strong emphasis in multiple regression textbooks aimed at economics and sociology, but popular regression textbooks in the behavioral sciences related to psychology and education do not express this same concern. For example, neither Cohen & Cohen (1975) nor Pedhazur (1982) deal with weighted least squares or logistic regression, two methods mentioned by multiple regression critics as preferable with dichotomous dependent variables. Both texts state that multiple regression can be used for and is mathematically equivalent to discriminant analysis when the dependent variable is a dichotomy (Cohen & Cohen, p. 442; Pedhazur, p. 687), but neither gives an indication that there are criticisms of this use. Tatsuoka (1971) states that in the dichotomous dependent variable case, multiple regression, discriminant analysis and canonical correlation are all mathematically equivalent and again, no indication is given of any criticisms of this approach.

Neter et al., (1983) list three problems that arise when the dependent variable is dichotomous: 1) non-normal error terms, 2) non-constant error variance, and 3) constraints on the response function. They state that even with binary dependent variables, ordinary least squares still provides unbiased estimators under quite general conditions, and "when the sample size is large, inferences concerning the regression coefficients and mean responses can be made in the same fashion as when the error terms are assumed to be normally distributed" (p. 357). They add, however, that these estimators will not be efficient, giving larger variances than could be obtained with weighted procedures.

The solutions proposed to these problems include using weighted least squares to give constant error variance and using a transformation (such as logistic) that limits the response function to a range of 0 to 1.

In comparing the use of logistic regression or discriminant analysis with dichotomous dependent variables, Press and Wilson (1978) suggest that logistic regression is preferred except when the populations are normal with identical covariance matrices. They extend the criticisms of others to include situations in which dichotomous variables are used as independent variables. They state that logistic regression is valid for a wide variety of underlying assumptions including 1) all explanatory variables are multivariate normally distributed with equal covariance matrices, 2) all explanatory variables are independent and dichotomous, and 3) some are multivariate normal and some dichotomous whereas discriminant analysis is only valid under the first set of assumptions. These comments are not directed at multiple regression, but would apply in those situations where it is mathematically equivalent to discriminant analysis. Their conclusion is that logistic regression with maximum liklihood estimation is preferred to linear discriminant analysis. They state, however, that it is unlikely that the two methods will give markedly different results or yield substantially different linear functions unless there is a large proportion of observations whose x-values lie in regions of the factor space with linear logistic response probabilities near zero or one. They go on to say that logistic regression is preferred when the normality assumptions are violated, especially when many of the independent variables are qualitative.

The critics state that in addition to the predictions made by the regression equation with a dichotomous dependent variable, statistical tests are also invalid. This would include the F test of the overall model and the t values for each predictor in the model.

Cox (1970), in referring to the use of multiple regression with dichotomous dependent variables, states that "the use of a model, the nature of whose limitations can be foreseen, is not wise, except for very limited purposes" (p. 18). If these critics are correct, it appears as if researchers in education and psychology should discontinue the use of multiple regression in these situations.

Problem

This paper is an attempt to assess the meaning of the charges made against multiple regression and to suggest what the regression community in education and psychology can do to come to terms with critics of multiple regression. The purpose of this paper is not to evaluate the validity of the criticisms but to deal with some logical extensions of them. If these criticisms are valid, are t-tests, analysis of variance, analysis of covariance, discriminant analysis, canonical correlation, and any use of dummy variables in multiple regression also called into question?

The questions raised by this paper, then, are:

1. To what extent do these criticisms affect the validity of other comparable statistical procedures?
2. If other statistical procedures using different assumptions give identical results to multiple regression using dichotomous dependent variables, does this imply suspicion concerning the other procedures or suspicion concerning the validity of the criticisms or both?

Procedures and Findings

To examine the validity and/or seriousness of these criticisms, implications of this situation are considered by examining a set of data taken from the A3 data set in Gunst & Mason (1980). This data set has 13 yearly observations with 14 variables. The year variable was dichotomized by letting the first 7 years be in one group and the last 6 years be the other group. The data is analyzed in 5 different cases with different arrangements of the dichotomous variable with one or two quantitative variables from this data set. The dichotomous variable is considered as both a dependent variable and an independent variable.

In Table 1 different combinations of quantitative and dichotomous independent and dependent variables where multiple regression has been used are presented with a listing of conventional alternative statistical methods and methods recommended by multiple regression critics. The critics suggest that in cases where a dichotomous dependent variable is used (cases 1 and 3) multiple regression is inappropriate. The approach taken in this paper is to compare the results of multiple regression in these cases with results of cases where multiple regression has not been attacked (cases 2 and 4).

Table 1

Possible Statistical Procedures to use with Different Combinations of Dichotomous and Quantitative Variables

Case	Dependent Variable	Independent Variable	Possible procedures
One Predictor			
1.	1 Dichotomous	1 Quantitative	Logistic regression Pearson correlation Pt. bis. correlation
2.	1 Quantitative	1 Dichotomous	t test Pearson correlation Pt. bis. correlation
Two+ Predictors			
3.	1 Dichotomous	2+ Quantitative/0+ Dichotomous	Logistic regression Discriminant analysis Multiple regression
4.	1 Quantitative	1+ Quantitative/1+ Dichotomous	Analysis of Covariance Multiple regression

Table 2 presents the results of the one predictor cases with the dichotomous variable as a dependent variable (case 1) and as an independent variable (case 2). In these situations the t value is the same whether the dichotomous variable is the independent or dependent variable. A one predictor model is the simplest case of multiple regression and the test of significance of the relationship is mathematically identical to an independent means t-test and a one-way ANOVA with two groups and the regression test of significance (t value) is the same whether the dichotomous variable is the independent or dependent variable. If a test of significance with a dichotomous dependent variable is invalid, then all tests of significance for an independent means t-test, a two-group one-way ANOVA and correlation/regression with an independent dichotomous variable are also invalid.

Table 2

One Predictor Examples

CASE 1: Multiple regression claimed to be invalid

Dependent variable = 2 (Dichotomous)
Independent variable = 3 (Quantitative)

t_3 = -6.910 -- same as case 2

CASE 2: Multiple regression is valid

Dependent variable = 3 (Quantitative)
Independent variable = 2 (Dichotomous)

t_2 = -6.910 -- same as case 1

Table 3 presents the results of the two predictor cases with the dichotomous variable as a dependent variable (case 3) and as an independent variable (cases 4a and 4b). Case 3 is a situation where multiple regression and discriminant analysis are both frequently used but is considered to be invalid by the critics of ordinary least squares due to the presence of a dichotomous dependent variable. The t values in case 3 are testing the significance of the relationship of each quantitative predictor with the dichotomous dependent variable controlled for the other quantitative predictor. Cases 4a and 4b give identical t values to those found in case 3 for the relationship between the dichotomous variable (which is now one of the

Table 3

Two Predictor Examples

CASE 3: Multiple regression claimed to be invalid

Dependent Variable = 2 (Dichotomous)
Independent Variables = 4 (Quantitative)
= 3 (Quantitative)

t_4 = -0.124 -- same as case 4a
t_3 = -6.480 -- same as case 4b

CASE 4: Multiple regression is valid

a. Dependent Variable = 4 (Quantitative)
Independent Variables = 2 (Dichotomous)
= 3 (Quantitative)

t_2 = -0.124 -- same as case 3
t_3 = -0.397

b. Dependent Variable = 3 (Quantitative)
Independent Variables = 4 (Quantitative)
= 2 (Dichotomous)

t_4 = -0.397
t_2 = -6.480 -- same as case 3

independent variables and in a legitimate place according to assumptions of multiple regression) and the dependent quantitative variable. If the tests for which the t values in Case 3 are invalid, then the tests for which the t values in cases 4a and 4b are used are also invalid. The t values in cases 4a and 4b are the same as the square root of the F values that would be computed with a one-way analysis of covariance in which the independent quantitative variable was treated as the covariate and the independent dichotomous variable as the grouping variable. So therefore if Case 3 is invalid, then all one-way ANCOVA designs and any use of dummy variables in multiple regression would be invalid also.

Conclusion and Recommendations

It is clear from the above examples that the tests of significance are identical whether the dichotomous variable is an independent variable or a dependent variable. It appears, therefore, that if the critics of using multiple regression with a dichotomous dependent variable are to be taken seriously, they must also deal with all significance testing with t tests, analysis of variance, analysis of covariance, discriminant analysis, and any use of dummy variables in multiple regression. There may be other statistics reported in a multiple regression analysis, such as the standard error of estimate or predicted values for which the interpretations may not be appropriate when dichotomous dependent variables are used, but this paper will not deal with these issues.

BIBLIOGRAPHY

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

Cox, D. R. (1970). The Analysis of Binary Data. London: Methuen & Co.

Goodman, L. A. (1978). Analyzing Qualitative/Categorical Data. Lanham, Maryland: University Press of America.

Cohen, J. & Cohen, P. (1975). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Gunst, R. F. & Mason, R. L. (1980). Regression Analysis and its Application. New York: Marcel Dekker.

Neter, J. et al. (1983). Applied Linear Regression Models. Homewood, Illinois: Richard D. Irwin.

Pedhazur, E. J. (1982). Multiple Regression in Behavioral Research. New York: Holt, Rinehart, Winston.

Press, S. J. & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, 73, 699-705.

Tatsuoka, M. M. (1971). Multivariate Analysis: Techniques for Educational and Psychological Analysis. New York: John Wiley & Sons.