KM 31(1989) pag 97 - 111

DETERMINING THE EFFECT OF A COMPOUND OF DUMMY VARIABLES OR POLYNOMIAL TERMS

Rob Eisinga, Peer Scheepers and Leo van Snippenburg *

Abstract

This paper discusses a method to obtain the standardized regression coefficient for a composite variable made up of dummy variables or polynomial terms. The method to be described enables the researcher to compare the effect of the composite variable with the effect of other predictor variables. Forming a composite variable is particularly useful in polynomial regression where individual regression coefficients are hard to interpret. A second type of application is assessing the impact of a compound of dummy variables. An empirical example dealing with the curvilinear relationship between church involvement and prejudice is used to illustrate the approach.

1 Introduction

Linear regression is one of the most frequently used statistical methods in the social sciences. However, the assumption of linear regression that the relationships between variables conform to a linear equation is highly restrictive to much social science research. In social science practice, relationships often depart from linearity and in those cases the application of ordinary regression is unwarranted. Two alternatives to linear regression which are useful when the assumption of linearity does not hold are polynomial regression and dummy variable regression. While considerable attention has been directed to the application of

* Department of Sociology, Catholic University of Nijmegen, P.O. Box 9108, 6500 HK Nijmegen, The Netherlands, phone: 080-515722/ 512377/515693. The authors would like to thank Jan Lammers for his suggestions and comments on a previous draft of this paper.

linear regression, the possibility of analyzing nonlinear relationships via polynomial regression or dummy variable regression has largely been neglected. The infrequent use of these models is unfortunate, but also understandable. Relatively little effort is required to estimate the regression coefficients for polynomial terms. Major difficulties arise, however, in interpreting these parameters. Likewise, the procedure for estimating the regression coefficients for individual dummy variables is straightforward. However, there is as yet no method ordinarily available to estimate the combined effect of a set of dummy variables.

This paper addresses both problems. Its purpose is to present a method which determines the standardized regression coefficient for a composite variable made up of dummy variables or polynomial terms. The method outlined below can be used when dealing with linear and a variety of nonlinear relationships. The example presented here confines itself to a specific form that nonlinearity may assume, i.e., the parabolic form.

To illustrate the method presented below, data were taken from the national survey "Social and cultural developments in the Netherlands", which was conducted in the autumn of 1985 (See Felling, et al., 1987). In the scientific study of religion it has been hypothesized that the relationship between church involvement and prejudice towards ethnic minorities is curvilinear, rather than linear. Prejudice, it has been postulated, increases as church involvement increases, but only to a certain point, after which a decline occurs. Another well-known predictor of prejudice is age. It is generally acknowledged that people become more prejudiced as they grow older. It is important to note that the interval variable labelled 'prejudice' contains standardized factor scores ranging from 0 to 1000, with the mean set at 500 and the standard deviation at 100. The variable 'church involvement' contains the following four categories: nonmembers, marginal church members, modal church members, and core church members. Of course, strictly speaking church involvement should not be considered an interval variable. Nevertheless, for the sake of simplicity both polynomial regression and dummy variable regression were carried out on the same sample data. Therefore, in

the illustration of polynomial regression church involvement was treated as interval-scaled.

2 Polynomial regression and dummy variable regression

Scientists generally agree that the relationship between variables is often nonlinear. Obviously, there is a large number of different forms a nonlinear relationship can take. Some of these relationships can be dealt with by polynomial equations. The general functional form of the polynomial equation is

$$Y = a + b_1X + b_2X^2 + ... + b_{k-1}X^{k-1} + e$$

where Y is the dependent variable, X the independent variable, (a) the intercept, b the unstandardized regression coefficient, and e the error term. In the polynomial equation, the independent variable X is raised to a certain power. The highest order to which the independent variable is raised indicates the degree of the polynomial. The highest order that the polynomial may take is equal to k-1, where k is the number of categories of the independent variable X, although a lower degree equation may often give a reasonably good fit to the data. The 4-category variable church involvement, for example, may be raised to the third power. In this case, the polynomial equation will yield predicted Y values that are equal to the means of the different Y arrays, thus resulting in the smallest possible value for the residual sum of squares.

Power polynomials can be dealt with by ordinary least squares regression, provided the variables are redefined and the nonlinear equation is converted into standard regression form by the appropriate transformation. To illustrate, consider the third-degree polynomial

$$Y = a + b_1X + b_2X^2 + b_3X^3 + e \qquad (1)$$

Because the original equation is difficult to deal with by means of ordinary least squares, we define two new variables and

substitute them into (1), in order to transform the nonlinear equation into linear form. If, in (1), we let

$$Z = X^2$$
$$\text{and } K = X^3$$

then the polynomial model becomes the familiar linear regression of Y on X, Z and K. Hence, power polynomials are linearizable by a suitable transformation and thereby amenable to ordinary least squares regression.

Another way of dealing with nonlinear relationships, particularly useful when the independent variables are discrete, is using dummy variables regression. In dummy variable regression, we let k-1 dummy variables represent the k categories of the original independent variable. When the k-1 dummy variables are employed as a set of independent variables predicting the dependent variable Y, the following equation results

$$Y = a + b_1D_1 + b_2D_2 + \ldots + b_{k-1}D_{k-1} + e$$

For example, to examine the relationship between church involvement and prejudice, the 4-category variable church involvement was broken down into the three dummy variables: D_1, D_2, and D_3. This breakdown was accomplished following the coding scheme given in Table 1.

Table 1. Breakdown of church involvement into dummy variables

	dummy variables		
	D_1	D_2	D_3
core church members	0	0	1
modal church members	0	1	0
marginal church members	1	0	0
nonmembers	0	0	0

When the dummy variables are used as independent variables in a regression analysis, the equation is given by

$$Y = a + b_1D_1 + b_2D_2 + b_3D_3 + e \qquad (2)$$

The regression coefficients in equation (2) have to be interpreted as follows. The intercept (a) represents the mean prejudice score of the reference category nonmembers. The unstandardized regression coefficients b_1, b_2, and b_3 represent the difference in mean score between nonmembers and marginal church members, nonmembers and modal church members, and nonmembers and core church members, respectively. According to equation (2), the relationship between church involvement and prejudice is not necessarily linear. The category means may occur in any pattern.
To illustrate polynomial and dummy variable regression, the third-degree equation (1) and the dummy variable equation (2) were applied to the data. However, because the proportion of variance incremented by the cubic term over and above the quadratic term was not statistically significant at the .05 level, the second-degree polynomial was considered more appropriate to describe the data than the third-degree polynomial. The dummy variable equation as well as the second-degree polynomial are given below. The results of the analyses are presented in Table 2.

Regression equations:

$$Y = a + b_1D_1 + b_2D_2 + b_3D_3 + e \qquad (2)$$
$$Y = a + b_1X + b_2X^2 + e \qquad (3)$$

where: Y = prejudice, D_1 = marginal church members, D_2 = modal church members, D_3 = core church members, X = church involvement, X^2 = church involvement squared, a = intercept, e = error term.

Table 2. Polynomial regression and dummy variable regression of prejudice on church involvement (a = intercept, b = unstandardized regression coefficient, and t = t-value, N=1566)

equation	variable	a	b	t
(2)	D_1	486	22.3	3.31
	D_2		35.9	5.49
	D_3		15.0*	1.84
(3)	X	438	58.2	4.42
	X^2		-10.4	-3.73

* = coefficient is not significant at .05 level.

According to the unstandardized regression coefficients for the dummy variables in Table 2, nonmembers are less prejudiced when compared to marginal church members and modal church members. We also find that modal church members have a higher mean score than both marginal church members and core church members. The difference between nonmembers and core church members is not statistically significant. These findings support the argument that the relationship between church involvement and prejudice is curvilinear, rather than linear. Table 2 also reveals that the unstandardized regression coefficient for X^2 in the second-degree polynomial is statistically significant at the .05 level ($t > 1.96$). This result indicates, once again, that the relationship between church involvement and prejudice is not linear, but parabolic.

To determine whether the deviations from linearity are statistically significant, we compared the proportion of variance accounted for by regression equation (2) and regression equation (3), with the proportion of variance accounted for by the linear equation $Y = a + bX + e$. The observed F ratios reveal that the variance accounted for by the linear equation is significantly lower than the variance accounted for by both regression equation (2) and regression equation (3) (See Krishnan Namboodiri, Carter and Blalock, 1975). Hence, we decided that the relationship between church involvement and prejudice statistically deviates from linearity.

3 The standardized solution

The estimation of the regression coefficients in the second-degree polynomial (3) is quite straightforward. However, polynomial regression analysis yields parameters that are not readily interpretable. The usual interpretation of the unstandardized regression coefficient as the change in Y associated with a one-unit change in X, controlling for the other independent variables, does not make sense in polynomial regression, because it is impossible for X to change its value while its powers are held constant. In polynomial regression, neither the coefficient for X nor the coefficients for the higher order terms can be interpreted

separately. In the second-degree polynomial, for instance, both X and X^2 have to be considered simultaneously.

However, there is another topic important to the interpretation of polynomial regression. The method of least squares depends on the calculation of the inverse of the correlation matrix. It is well-known that computational difficulties arise if the correlation matrix is singular or ill-conditioned. Ill-conditioned data occur if the correlation between the independent variables is near unity. The consequences of this situation, which is referred to as collinearity, can be severe. In particular, the unstandardized regression coefficients tend to be "inflated" so that predicted Y values may be unreasonable. Moreover, the standardized regression coefficients may exceed unity and have an incorrect sign. As collinearity increases, the standard errors for the regression coefficients tend to become larger and the confidence intervals tend to become wider (e.g., Farrar and Glauber, 1967; Mason, Gunst and Webster, 1975).

In polynomial regression, collinearity of the predictor variables is, in a sense, self-induced. Powered terms, especially when they are made up of positive values, tend to be highly correlated.

The question what should be done with collinearity in polynomal regression does not have a simple answer. One prescription, recommended by several authors, is substracting the mean from the independent variable X (e.g., Marquardt and Snee, 1975; Cohen and Cohen, 1975: 227; Opp and Schmidt, 1976: 198-199; Bradley and Srivastava, 1979). Centering X attenuates the correlation between X and its powers, and thereby reduces the inflation of the unstandardized regression coefficients.

Centering X leaves the coefficient of determination and the tests for statistical significance unaffected. This should not be surprising. It refers to the property of the method of least squares called scale invariance, indicating that if any of the independent variables are scaled by addition of a constant or by multiplication by a constant, scale-free quantities such as R^2 and test statistics (t and F-values) will remain unchanged.

In the linear equation, both the standardized and the unstandardized regression coefficients are also invariant under centering. This pleasant property, however, does not apply to power polynomials. To be sure, substracting the mean from X has no

effect on the unstandardized regression coefficient for the highest order term, for instance, X^2 in the second-degree polynomial. However, as we will see, in the second-degree equation, centering X causes both the unstandardized and the standardized regression coefficient for X, and the standardized regression coefficients for X^2 to change (e.g., Cohen, 1978; Jagodzinski and Weede, 1980: 141; Pedhazur, 1982: 414). This illustrates, once again, that in polynomial regression, the regression coefficients do not lend themselves to clear-cut interpretations.

As indicated earlier, because it is impossible to conceive the unstandardized regression coefficients in the polynomial equation as expressing the effect of one regressor, while the others are fixed, X and its powers have to be considered simultaneously. Therefore, it may be desirable to find some measure for the effect of X and the higher order terms considered as a single variable, but in practice left as a set of distinct regressors. Thus, our intention is to obtain the effect on Y for X and its powers taken together. The method to be explicated here has occasionally been proposed by Coleman (1976), and Jagodzinski and Weede (1980: 141; 1981). This paper, however, clarifies the key statements and extends the approach.

In order to obtain the combined effect of X and its powers, a composite variable T is defined, that replaces X and the higher order terms. This composite variable T is computed as the weighted sum of X and its powers, using the previously estimated unstandardized regression coefficients for X and the higher order terms as weights. This composite variable is subsequently used in a second regression run. To illustrate, if we call the parenthetic component in

$$Y = a + b(b_1X + b_2X^2 + ... + b_{k-1}X^{k-1}) + e \qquad (4)$$

T, equation (4) simplifies to

$$Y = a + bT + e \qquad (5)$$

where T represents the composite polynomial and b the unstandardized regression coefficient for T. The regression of Y

on T is identical to the regression of Y on X and its powers, with respect to the intercept (a) and the proportion of variance accounted for. Furthermore, as we will show, in designs including predictor variables Z_j linearly related to Y, as in

$$Y = a + b(\sum_{k=1}^{K} b_{k-1} X^{k-1}) + \sum_{j=1}^{J} b_j Z_j + e$$

estimating the parameters afresh in a second regression run has no effect on both the unstandardized and the standardized regression coefficients for Z_j.

It should be pointed out that the unstandardized regression coefficient for T in equation (5) is not identical to the standardized regression coefficient, as Coleman (1976: 15) suggested, but, of course, always equals 1. It is also important to note that the standardized regression coefficient for T differs from an ordinary standardized regression coefficient. Usually, the standardized regression coefficient (β) for X is equal to

$$\beta = b \cdot \sigma X / \sigma Y$$

Because the unstandardized regression coefficient for T is equal to 1, however, the standardized regression coefficient for T is simply the standard error of T (σT) divided by the standard error of Y (σY). This implies that the standardized regression coefficient for T will always have a positive value. Consequently, the sign of the standardized regression coefficient for T is a technical artifice.

Recall that centering X in power polynomials affects the standardized regression coefficients for X and the higher order terms. The standardized regression coefficient for T, however, remains unchanged under linear transformation. Hence, this coefficient can be interpreted as the effect of T with respect to the effect of other predictor variables.

Let us now turn to dummy variable regression. When dummy variables are used in a regression analysis, the result is a set of regression coefficients for many individual dummy variables. Dummy variable regression provides measures for the relationship between

one aspect of the original independent variable X and the dependent variable Y. Therefore, it might be useful to obtain a measure for the effect of all the dummy variables taken together. This effect may then be compared with the effect of other independent variables.

The construction of the composite variable can, again, be accomplished by defining a new variable from the weighted sum of the distinct dummy variables. If we call the linear combination $(b_1D_1 + b_2D_2 + ... + b_{k-1}D_{k-1})$ composite variable T, the new regression equation is given by

$$Y = a + b(b_1D_1 + b_2D_2 + ... + b_{k-1}D_{k-1}) + e = a + bT + e$$

Again, the regression of Y on T is identical to the regression of Y on the original dummy variables, in the intercept (a) and in the proportion of variance accounted for.

In order to fully explain the procedure outlined above, the variable age (Z), which is linearly related to prejudice, was added to the earlier reported regression equations (2) and (3). Six regression analyses were applied to the data. The equations are listed below. The regression summaries are given in Table 3. To illustrate the effect of centering, in equation (8) and equation (9) the mean was substracted from X prior to the squaring operation.

Regression equations:

$$Y = a + b_1X + b_2X^2 + b_3Z + e \qquad (6)$$

$$Y = a + b_4(b_1X + b_2X^2) + b_3Z + e = a + b_4T + b_3Z + e \qquad (7)$$

$$Y = a + b_1(X - \bar{X}) + b_2(X - \bar{X})^2 + b_3Z + e \qquad (8)$$

$$Y = a + b_4(b_1(X - \bar{X}) + b_2(X - \bar{X})^2) + b_3Z + e = a + b_4T + b_3Z + e \qquad (9)$$

$$Y = a + b_1D_1 + b_2D_2 + b_3D_3 + b_4Z + e \qquad (10)$$

$$Y = a + b_5(b_1D_1 + b_2D_2 + b_3D_3) + b_4Z + e = a + b_5T + b_4Z + e \qquad (11)$$

where: $\underline{Y}$ = prejudice, X = church involvement,
$\bar{X}$ = church involvement mean,
X^2 = church involvement squared,
Z = age, D_1 = marginal church members,
D_2 = modal church members, D_3 = core church members,
T = weighted sum of polynomial terms or dummy variables,
a = intercept, e = error term.

Table 3. Polynomial regression and dummy variable regression of prejudice on church involvement and age (a = intercept, b = unstandardized regression coefficient, beta = standardized regression coefficient, t = t-value, and R^2 = proportion of explained variance, N=1566)

equation	variable	a	b	beta	t	R^2
(6)	X	361	52.3	.56	4.16	.10936
	X^2		-10.4	-.53	-3.90	
	Z		2.2	.31	12.51	
(7)	T	361	1.0	.10	4.30	.10936
	Z		2.2	.31	12.74	
(8)	$(X - \bar{X})$	424	10.5	.11	3.75	.10936
	$(X - \bar{X})^2$		-10.4	-.12	-3.90	
	Z		2.2	.31	12.51	
(9)	T	424	1.0	.10	4.30	.10936
	Z		2.2	.31	12.74	
(10)	D_1	403	19.4	.08	3.02	.10943
	D_2		22.1	.09	3.50	
	D_3		-0.7	-.00*	-0.09	
	Z		2.2	.30	12.47	
(11)	T	403	1.0	.10	4.31	.10943
	Z		2.2	.30	12.71	

* = coefficient is not significant at .05 level

What can we conclude with respect to the polynomial regressions of prejudice on church involvement and age? Well, first of all, from equation (6) and equation (8) we can conclude that substracting the mean from X affects both the unstandardized and the standardized regression coefficients for X, as well as the standardized regression coefficient for X^2. However, the unstandardized and the standardized regression coefficients for Z, the unstandardized regression coefficient for X^2, and the t-values for Z and X^2, remain unchanged under centering. Further, centering X prior to the squaring operation has no effect on the t-value and the standardized regression coefficient for the weighted sum of X and X^2. The standardized regression coefficient for T indicates

the effect of the curvilinear predictor church involvement with respect to the effect of the variable age (Z).

Second, the regression summaries indicate that running regression with the composite variable T has no effect whatsoever on the intercept (a) and the (rather low) proportion of variance accounted for (R^2). The intercept and the coefficient of determination for equation (6) and equation (7), as well as for equation (8) and equation (9), correspond. In passing, it might be noted that the composite variable T in equation (7) has been derived from regression equation (6) and not from regression equation (3). The relative weights of X and X^2 should not only be determined by the polynomial terms themselves, but also by the other independent variables in the regression equation, in our example, the variable age (Z) (e.g., Igra, 1979; Jagodzinski and Weede, 1981).

And what can we conclude with respect to the dummy variable regressions of prejudice on church involvement and age? First of all, from equation (10) we can conclude that nonmembers are still less prejudiced than both marginal church members and modal church members. The difference in prejudice between nonmembers and core church members is not statistically significant.

Second, as in polynomial regression, the intercept (a) and the proportion of variance accounted for by regression equation (11), remain as they were in equation (10). The proportion of variance explained by the dummy variables is somewhat higher than the proportions of variance explained by the second-degree polynomial because the former incorporates more predictor variables. Further, comparison of the results for equation (10) with the results for equation (11) demonstrates, once more, that forming a composite variable has no effect on the regression coefficients for the variable not belonging to the composite, that is, the variable age (Z).

Last, the standardized regression coefficient for T assesses the effect of the dummy variable set on prejudice. This effect can be compared with the effect of Z in equation (11). Table 3 shows that the variable age (Z) is more important to prejudice than the composite variable T.

A final point should be made regarding the interpretation of the regression coefficients for the composite variable T. The unstandardized regression coefficient b for T does not permit meaningful interpretation, because b can be manipulated almost at will. To explain this we have to recall that the composite variabele T in the second-degree polynomial, for instance, is a weighted linear combination of X and X^2. The weights b_1 and b_2 represent, in fact, the ratio of the effects of X and X^2. Running regression with any linear transformation of these weights, will yield exactly the same intercept, coefficient of determination, and standardized regression coefficient for T. However, the unstandardized regression coefficient for T is sensitive to linear transformation. Hence, the solution for b is not unique.

With respect to the standardized regression coefficient for T, we have to point out that in our example this coefficient indicates the effect of a nonmonotonic predictor variable T on Y. It represents the combined effect of the set of discrete dummy variables or polynomial terms. The actual form of the relationship between X and Y, however, is described by the unstandardized regression coefficients for the individual dummy variables or polynomial terms. In dummy variable regression, the unstandardized regression coefficients indicate the difference in mean scores between any particular category of X and the reference category. In polynomial regression, differential calculus may be used to obtain the minima and the maxima of the polynomial. The derivate of the polynomial equation provides the unstandardized conditional effect of X on Y at any particular value of X, for instance, the value of X at which the curve bends.

4 Conclusions

The purpose of this paper was to present a method to obtain the standardized regression coefficient for a composite variable made up of dummy variables or polynomial terms. In closing, we iterate the suggested procedure. After first carrying out a full regression on the dummy variables or polynomial terms, along with other predictor variables, a composite variable is created, using the previously estimated regression coefficients for the dummy

variables or the polynomial terms as weights. This newly-defined variable replaces the dummy variables or the polynomial terms in a second regression run. The results for the second regression are identical to the results for the first regression with respect to the intercept, the proportion of variance accounted for, and the regression coefficients for the variables not belonging to the composite variable. The standardized regression coefficient for the composite variable reveals the effect of the composite variable with respect to the effect of the other predictor variables in the equation.

References

Bradley, R.A. and Srivastava, S.S. (1979). "Correlation in polynomial regression", American Statistician 33: 11-14.

Cohen, J. (1978). "Partialed products are interactions; partialed powers are curve components", Psychological Bulletin 85: 858-866.

Cohen, J. and Cohen, P. (1975). Applied multiple regression/ correlation analysis for the behavioral sciences, New York: John Wiley and Sons.

Coleman, J.S. (1976). "Regression analysis for the comparison of school and home effects", Social Science Research 5: 1-20.

Farrar, D.E. and Glauber, R.G. (1967). "Multicollinearity in regression analysis: the problem revisited", Economics and Statistics 49: 92-107.

Felling, A., Peters, J., Schreuder, O., Eisinga, R. and Scheepers, P. (1987). Religion in Dutch society 85. Documentation of a national survey on religious and secular attitudes in 1985. Amsterdam: Steinmetz Archive.

Igra, A. (1979). "On forming variable set composites to summarize a block recursive model", Social Science Research 8: 253-264.

Jagodzinski, W. and Weede, E. (1980). "Weltpolitische und ökonomische Determinanten einer ungleichen Einkommensverteilung: Eine international vergleichende Studie", Zeitschrift für Soziologie 9: 132-148.

Jagodzinski, W. and Weede, E. (1981). "Testing curvilinear propositions by polynomial regression with particular references to the interpretation of standardized solutions", Quality and Quantity 15: 447-463.

Krishnan Namboodiri, N., Carter, L.F. and Blalock, H.M.jr. (1975). Applied multivariate analysis and experimental designs. New York: McGraw-Hill.

Marquardt, D.W. and Snee, R.B. (1975). "Ridge regression in practice", American Statistician 29: 3-20.

Mason, R.L., Gunst, R.F. and Webster, J.T. (1975). "Regression analysis and problems of multicollinearity", Communications in Statistics 4: 277-292.

Opp, K.D. and Schmidt, P. (1976). Einführung in die Mehrvariabelenanalyse. Reinbek bei Hamburg: Rohwolt.

Pedhazur, E.J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Ontvangen: 14-06-1988
Geaccepteerd: 20-04-1989