International Journal of Statistics and Systems
ISSN 0973-2675 Volume 11, Number 2 (2016), pp. 135-143

http://www.ripublication.com

Regression Analysis with Categorical Variables

M. Venkataramana [1], Dr. M. Subbarayudu[2], M. Rajani[3], Dr. K.N. Sreenivasulu[4]

[1,3] Research Scholar Department of Statistics, S.V. University, Tirupati-517502, AP, India.
E-mail: ramana.astat@yahoo.in, rajani231190@gmail.com.

[2]Correspondence Author, Professor and Head, Department of Statistics, S.V. University, Tirupati-517502, AP, India.
E-mail: msrsvu@yahoo.co.in

[4]Assistant Professor & Head, Department of Statistics, ANGR. Ag. University, Ag. College, Mahanandi, Kurnool (Dt), India.
E-mail: kotha.sreenu@gmail.com

Abstract

In regression analysis the dependent variable is frequently influenced not only by the variables that can be readily quantified on some well defined scale but also by the variables that are essentially qualitative in nature .Qualitative variables with different coding methods make the linear regression model an extremely flexible tool, that is capable of handling many interesting problems encountered in empirical problems. Qualitative variables often called Categorical variables, dummy variables indicator variables. The purpose of this paper is to incorporate categorical independent variables into the regression model through both dummy and effect coding methods at a time and to analyze the data.

Keywords: ANOVA model, Categorical variables, Combination of dummy and effect coding.

INTRODUCTION

Qualitative variables usually indicate the presence or absence of a quality or an attribute such as male or female, black or white, catholic or non catholic, citizen or non citizen. A regression model which contains explanatory variables that are exclusively qualitative in nature is known as analysis of variance (ANOVA) model and a model which contains the explanatory variables as an admixture of both quantitative and qualitative variables is known as analysis of covariance (ANCOVA) model.

The use of categorical variables as independent variables in the regression model involves the application of coding methods. Coding methods refer to ways in which membership in a given group can be represented in mutually exclusive and exhaustive manner. Any qualitative variable with k categories or classes can be represented by creating (k-1) dummy variables that takes on numerical values. This process involves assigning one numerical value which is called a code to all subjects of a particular group and a different numerical value to all those of the other groups to convert the qualitative variables into quantitative variables to run the regression.

DUMMY CODING METHOD

Dummy variables are used as a classifying device in that they divide the entire sample into various groups based on qualities and implicitly allow to run the individual regressions for each sub group. Dummy coding method represents a group membership with dummy variables that take on values 0 and 1. That is membership in a particular group is coded as one where as non membership in a group is coded as zero. However assignment of 1 and 0 values to categories is arbitrary.

The category to which the value zero is assigned is often referred as base, bench mark, control, comparison or omitted category. The number of dummy variables must be less than the number of categories or classifications of each qualitative variable to avoid dummy variable trap. The coefficients attached to the dummy variables must always be interpreted in relation to the base or reference group that assign value zero. If a model has several qualitative variables with several classes introduction of dummy variables can be consuming a large number of degrees of freedom. Dummy explanatory variables are denoted by the symbol 'D' rather than by the usual symbol 'X' in regression model to emphasize that we are dealing with qualitative variables.

The regression model involving one qualitative variable as independent variable with k categories or classes, using dummy coding can be represented as

$$Y_{ij} = B_0 + \sum_{j=1}^{k-1} B_j D_{ij} + \epsilon_{ij} \quad (1)$$

where

- Y_{ij} = The score on the dependent variable for subject i in group j
- B_0 = The intercept that represents the mean of the group coded 0 on all the dummy variables.
- K = number of categories or classifications of dummy independent variable.
- B_j = The regression coefficient associated with j^{th} group, it represents the difference between the mean of the group coded 1 on the corresponding dummy variable and the mean of the group coded 0 on all the dummy variables.
- D_{ij} = The numerical value of dummy variable assigned to subject i^{th} in the j^{th} group.
- ϵ_{ij} = The error associated with i^{th} subject in the j^{th} group.

EFFECT CODING METHOD

Effect coding is appropriate where each group is compared with entire set of groups rather than with a reference group. That is effect coding is useful in testing the effect of a treatment defined as the deviation between treatment mean and the grand mean. In effect coding the dummy variables takes the values 1, 0 and -1. The coding method used for effect coding is similar to that used for dummy coding except for the way in which the reference group is identified. Using dummy coding the reference group is coded as 0 but in the effect coding it is coded as -1.

A regression model involving one qualitative variable as independent variable with k categories or classes using effect coding can be represented as

$$Y_{ij} = C_0 + \sum_{j=1}^{k-1} C_j E_{ij} + \epsilon_{ij} \quad (2)$$

where

Y_{ij} = The score on the dependent variable Y for subject i in group j

C_0 = The intercept that represents the grand mean of the dependent variable for all groups.

k = number of categories of the dummy independent variable.

C_j = Regression coefficient associated with j^{th} group, it represents the difference between the mean of the group coded 1 on the corresponding dummy variable and the grand mean of all groups.

E_{ij} = The numerical value of dummy variable assigned to subject i^{th} in the j^{th} group.

ϵ_{ij} = The error associated with the i^{th} subject in the j^{th} group.

EXAMPLE FOR DUMMY CODING AND EFFECT CODING

Hussain Alkharusi (2012) explained the use of categorical variables in regression analysis through dummy and effect coding methods independently with an example. In dummy coding method it is observed that the test of significance of a given regression coefficient is equivalent to a test of difference between the mean of the group associated with the regression coefficient and the mean of the reference group. In effect coding method each group is compared with the entire set of groups rather than with a reference group. Here the test for significance of regression coefficient is equivalent to testing the significance of the treatment effect. The two methods, dummy coding and effect coding gives entirely different interpretations regarding regression coefficients. But in both the methods the values of R^2 is same.

In studying the relationship between student's exposure to different types of teaching methods and their performance, the students were divided into three groups consisting of five each which were taught by discovery method, observational method and traditional method. To run the regression with dummy coding two dummy variables D_1 *and* D_2 would be needed to classify the three groups. Similarly to run the regression with effect coding two dummy variables E_1 and E_2 are used to represent the information in the classification of teaching methods. Table 1 displays the dummy coding and effect coding of the independent variable teaching method. In dummy coding method the subjects in the discovery group have been coded as 1 for D_1 and 0

for D_2 , those in the observational group have been coded as 0 for D_1 *and as* 1 for D_2 and those in the traditional group have been coded as zero for both D_1 and D_2. As such the traditional group served as reference group. And in effect coding method scores of discovery group receive a 1 on E_1 and a 0 on E_2 , scores of observational group receive a 0 on E_1 and a 1 on E_2 and scores of traditional group receive values -1 on both $E_1$and E_2 .

Hence the regression equations through dummy coding and effect coding respectively given by

$$Y = B_0 + B_1 D_1 + B_2 D_2 + \varepsilon_1 \quad (3)$$

And $$Y = C_0 + C_1 E_1 + C_2 E_2 + \varepsilon_2 \quad (4)$$

Table 1: Dummy coding and effect coding for data of three groups

Group	score	Dummy coding		Effect coding	
		D_1	D_2	E_1	E_2
Discovery	90	1	0	1	0
	88	1	0	1	0
	91	1	0	1	0
	95	1	0	1	0
	93	1	0	1	0
Observational	78	0	1	0	1
	74	0	1	0	1
	74	0	1	0	1
	74	0	1	0	1
	70	0	1	0	1
Traditional	56	0	0	-1	-1
	59	0	0	-1	-1
	54	0	0	-1	-1
	55	0	0	-1	-1
	58	0	0	-1	-1

The estimated regression equation for dummy coded data is

$$\hat{Y} = 56.40 + 35.00\, D_1 + 17.40 D_2 \quad (5)$$

[1.2329] [1.7436] [1.7436]

(45.7464) (20.0739) (9.9796)

R^2=0.9711, F=201.4825

The intercept B_0 =56.40 represents mean of traditional group (reference group) since both the two dummy variables D_1 and D_2 are coded as 0 in this group. The regression coefficient of D_1 (B_1=35) represents mean of the discovery group is 35 greater than that of traditional group. And the regression coefficient of D_2 (B_2=17.40) represents the mean of observational group is 17.40 greater than that of traditional group. Since

both B_1 and B_2 are statistically significant we conclude that the difference between mean of the traditional group and mean of the discovery group is significant also the difference between mean of the traditional group and mean of the observational group are significant .Also $\mathbf{R^2}$ is statistically significant and conclude that the mean scores of three groups are different. No conclusion can be made regarding the difference between mean of the discovery group to that of observational group. This may be tested using multiple comparison test.

The estimated regression equation for effect coded data is

$$\hat{Y} = 73.8667 + 17.5333E_1 - 0.0667E_2 \qquad (6)$$

[0.7118] [1.0066] [1.0066]

(103.7737) (17.4175) (-0.0066)

R^2= 0.9711, F= 201.4825

The intercept C_0=73.8667 represents the grand mean if all groups. The regression coefficient of $E_1(C_1 = 17.5333)$ indicates that the mean of the discovery group is 17.5333 greater than the grand mean of all groups and this difference is statistically significant. Similarly the regression coefficient of $E_2(C_2 = -0.0667)$ indicates that the mean of observational group is 0.0667 smaller than the grand mean of all groups and this difference is not statistically significant. Also R^2=0.9711 is statistically significant.

The values in square brackets and in parenthesis of equations (5) and (6) respectively indicates standard errors and t-values.

COMBINATION OF DUMMY AND EFFECT CODING METHOD

In this connection, it is proposed to fit a single regression model to analyze the data through dummy coding and effect coding methods at a time instead of fitting two regressions one for each method. The procedure is explained through an example problem of student's scores through three teaching methods.

For this purpose we may club the two regression equations (3) and (4) into a single regression using dummy variable technique without loss of information and can be written as

$$Y = a_0 + a_1D + a_2(D_1 - DD_1) + a_3(D_2 - DD_2) + a_4DE_1 + a_5DE_2 + u_1 \qquad (7)$$

Put D = 0 for dummy coding

= 1 for effect coding

Alternatively equation (7) may be written as

$$Y = b_0 + b_1D + b_2DD_1 + b_3DD_2 + b_4(E_1 - DE_1) + b_5(E_2 - DE_2) + u_2 \qquad (7')$$

Put D = 1 for dummy coding

= 0 for effect coding

Either from the estimated equation (7) or (7') we can easily derive the actual values of intercept and regression coefficients for both the dummy and the effect coding methods as follows:

Hence from the estimated equation (7) the estimated equations through dummy coding and effect coding methods respectively given as

$\widehat{Y} = a_0 + a_2 D_1 + a_3 D_2$ (8)

And $\widehat{Y} = (a_0 + a_1) + a_4 E_1 + a_5 E_2$ (9)

Similarly form the estimated equation (7') the estimated regression equations through dummy coding and effect coding methods respectively given as

$\widehat{Y} = (b_0 + b_1) + b_2 D_1 + b_3 D_2$ (8')

And $\widehat{Y} = b_0 + b_4 E_1 + b_5 E_2$ (9')

Note that the pair of equations (8) & (8') and (9) & (9') gives identical results.

Example

The proposed method is applied on the same example problem of student's exposure to three teaching methods and their performance score. To run the regression equations (7) and (7') the corresponding data may be considered in table (2) and in table (3) respectively. The columns (1) – (5) in these two tables are the repetition of columns of table (1) twice one below the other. Columns (6) is data on dummy variable D which divides the entire data into two groups namely dummy coding method and effect coding method. The columns (7) – (10) are calculated accordingly.

Table 2:

Y (1)	D_1 (2)	D_2 (3)	E_1 (4)	E_2 (5)	D (6)	D_1 - DD_1 (7)	D_2 - DD_2 (8)	DE_1 (9)	DE_2 (10)
90	1	0	1	0	0	1	0	0	0
88	1	0	1	0	0	1	0	0	0
91	1	0	1	0	0	1	0	0	0
95	1	0	1	0	0	1	0	0	0
93	1	0	1	0	0	1	0	0	0
78	0	1	0	1	0	0	1	0	0
74	0	1	0	1	0	0	1	0	0
71	0	1	0	1	0	0	1	0	0
76	0	1	0	1	0	0	1	0	0
70	0	1	0	1	0	0	1	0	0
56	0	0	-1	-1	0	0	0	0	0
59	0	0	-1	-1	0	0	0	0	0
54	0	0	-1	-1	0	0	0	0	0
55	0	0	-1	-1	0	0	0	0	0
58	0	0	-1	-1	0	0	0	0	0

90	1	0	1	0	1	0	0	1	0
88	1	0	1	0	1	0	0	1	0
91	1	0	1	0	1	0	0	1	0
95	1	0	1	0	1	0	0	1	0
93	1	0	1	0	1	0	0	1	0
78	0	1	0	1	1	0	0	0	1
74	0	1	0	1	1	0	0	0	1
71	0	1	0	1	1	0	0	0	1
76	0	1	0	1	1	0	0	0	1
70	0	1	0	1	1	0	0	0	1
56	0	0	-1	-1	1	0	0	-1	-1
59	0	0	-1	-1	1	0	0	-1	-1
54	0	0	-1	-1	1	0	0	-1	-1
55	0	0	-1	-1	1	0	0	-1	-1
58	0	0	-1	-1	1	0	0	-1	-1

Table 3:

Y (1)	D_1 (2)	D_2 (3)	E_1 (4)	E_2 (5)	D (6)	D D_1 (7)	D D_2 (8)	E_1-DE_1 (9)	E_2-DE_2 (10)
90	1	0	1	0	0	0	0	1	0
88	1	0	1	0	0	0	0	1	0
91	1	0	1	0	0	0	0	1	0
95	1	0	1	0	0	0	0	1	0
93	1	0	1	0	0	0	0	1	0
78	0	1	0	1	0	0	0	0	1
74	0	1	0	1	0	0	0	0	1
71	0	1	0	1	0	0	0	0	1
76	0	1	0	1	0	0	0	0	1
70	0	1	0	1	0	0	0	0	1
56	0	0	-1	-1	0	0	0	-1	-1
59	0	0	-1	-1	0	0	0	-1	-1
54	0	0	-1	-1	0	0	0	-1	-1
55	0	0	-1	-1	0	0	0	-1	-1
58	0	0	-1	-1	0	0	0	-1	-1
90	1	0	1	0	1	1	0	0	0
88	1	0	1	0	1	1	0	0	0
91	1	0	1	0	1	1	0	0	0
95	1	0	1	0	1	1	0	0	0
93	1	0	1	0	1	1	0	0	0
78	0	1	0	1	1	0	1	0	0
74	0	1	0	1	1	0	1	0	0
71	0	1	0	1	1	0	1	0	0

76	0	1	0	1	1	0	1	0	0
70	0	1	0	1	1	0	1	0	0
56	0	0	-1	-1	1	0	0	0	0
59	0	0	-1	-1	1	0	0	0	0
54	0	0	-1	-1	1	0	0	0	0
55	0	0	-1	-1	1	0	0	0	0
58	0	0	-1	-1	1	0	0	0	0

The regression results of equation (7) are given by

$$\hat{Y} = 56.4 + 17.4667D + 35(D_1 - DD_1) + 17.4(D_2 - DD_2) + 17.5333\,DE_1 - 0.0667\,DE_2 \quad (10)$$

[1.2329] [1.4236] [1.7436] [1.7436] [1.0066] [1.0066]
(45.7464) (12.2693) (20.0739) (9.9796) (17.4176) (-0.0662)

$R^2 = 0.97108$, F=161.186

Putting D=0,1 in equation (10) we get the estimated equations through dummy coding, effect coding and presented respectively in equations (11) and (12) given below.

$$\hat{Y} = 56.4 + 35D_1 + 17.4D_2 \quad (11)$$

$$\hat{Y} = 73.8667 + 17.5333E_1 - 0.0667E_2 \quad (12)$$

Similarly the regression results of equation (7') are given by

$$\hat{Y} = 73.8667 - 17.4667D + 35\,DD_1 + 17.4\,DD_2 + 17.5333(E_1 - DE_1) - 0.0667(E_2 - DE_2) \quad (10')$$

[0.7118] [1.4236] [1.7435] [1.7435] [1.0066] [1.0066]
(103.7737) (-12.2693) (20.0739) (9.9796) (17.4176) (-0.0662)

$R^2 = 0.97108$, F=161.186

The value in square brackets and in parenthesis of equations (10) and (10') respectively indicates standard errors and t-calculated values.

Putting D=1,0 in equation (10') we get the estimated equations through dummy coding, effect coding and presented respectively in equations (11') and (12') given below

$$\hat{Y} = 56.4 + 35D_1 + 17.4D_2 \quad (11')$$

$$\hat{Y} = 73.8667 + 17.5333E_1 - 0.0667E_2 \quad (12')$$

It is observed that the pair of equations (11),(11') and (12),(12') are identical.

CONCLUSIONS

Hussain Alkharusi (2012) has described how categorical independent variables can be incorporated into regression model through dummy coding and effect coding separately. With an example, it is observed that, both the coding methods give identical R^2, but the two methods differ in the information provided by the regression equations. The purpose of this paper is to propose a method to incorporate categorical

independent variables into the regression model through both the dummy coding and effect coding at a time. In this context two alternative methods are proposed. It is observed from the empirical analysis that both the proposed methods give identical results.

In proposed method, we are running only one regression to analyze the data through dummy coding and effect coding at a time which takes less time and obtain the same results instead of running two regressions separately for dummy coding and effect coding. Hence we may conclude that the proposed method is preferable to Hussain Alkharusi (2012) method.

REFERENCE

[1] Damodar Gujarati (19760): Use of Dummy variables in testing for equality between sets of coefficients in linear regressions: A generalization, The American Statistician, Vol. 24, No. 5, pp: 18-22.

[2] Hussain Alkharusi (2012).Categorical variables in Regression Analysis: A comparison of Dummy and Effect coding, International Journal of Education, ISSN: 1948-5476, Vol.4, No.2, pp: 202-210.