Macrothink Institute™

International Journal of Education
ISSN 1948-5476
2012, Vol. 4, No. 2

Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding

Hussain Alkharusi*

**College of Education, Sultan Qaboos University*

P.O.Box: 32 Al-Khod, P.C.: 123, Sultanate of Oman

Tel: 968-9622-2535   E-mail: hussein5@squ.edu.om

Received: January 14, 2012   Accepted: March 18, 2012   Published: June 1, 2012

doi:10.5296/ije.v4i2.1962   URL: http://dx.doi.org/10.5296/ije.v4i2.1962

Abstract

The use of categorical variables in regression involves the application of coding methods. The purpose of this paper is to describe how categorical independent variables can be incorporated into regression by virtue of two coding methods: dummy and effect coding. The paper discusses the uses, interpretations, and underlying assumptions of each method. In general, overall results of the regression are unaffected by the methods used for coding the categorical independent variables. In any of the methods, the analysis tests whether group membership is related to the dependent variables. Both methods yield identical R^2 and F. However, the interpretations of the intercept and regression coefficients depend on what coding method has been applied and whether the groups have equal sample sizes.

Keywords: *categorical variables; regression analysis; coding methods; dummy coding; effect coding; dummy variables*

1. Introduction

The use of regression analysis requires that all variables entered into the model be continuous variables. A continuous variable is one on which subjects differ in amount or degree such as study time and height. However, it is possible to include categorical independent variables in the regression analysis. A categorical variable is one for which the units of observations differ in terms of type or kind such as a group membership(e.g., gender, marital status) or assignment to a treatment condition (e.g., experimental or control) (Allen, 1997; O'Grady & Medoff, 1988; Pedhazur, 1997).

The use of categorical independent variables in the regression analysis involves the application of coding methods. There are a number of coding methods (Cohen & Cohen, 1983). However, only two of the more common methods, dummy and effect coding, are discussed in this paper. The paper begins with an overview of the coding methods including general guidelines for coding categorical variables. Then, a description of the dummy and effect coding along with an example that illustrates their use and interpretation in the regression analysis are provided. Also, the paper addresses the use of these coding methods in designs with unequal sample sizes. Finally, the paper highlights the similarity between analysis of variance and multiple regression.

2. Overview of Coding Methods

Coding methods refer to ways in which membership in a group can be represented in a mutually exclusive and exhaustive manner. In general, any categorical variable with *k* categories can be represented by creating (*k*-1) dummy variables that take 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. This is because data need to be represented quantitatively for the purpose of regression analysis and that categorical variables lack this property (Keppel & Zedeck, 1989; O'Grady & Medoff, 1988; Cohen & Cohen, 1983).

3. Dummy Coding method

3.1 Definition

This method represents group membership with dummy variables that take on values 0 and 1. In other words, membership in a particular group is coded one whereas non-membership in the group is coded zero. In most common applications, one group receives 0s on all dummy variables and functions as the reference group (Cohen & Cohen, 1983; Myers & Well, 2003).

3.2 Structural Model

When dummy coding is used in the regression analysis, the overall results indicate whether there is a relationship between the dummy variables and the dependent variables. The values of the intercept and the regression coefficients of the resulted regression model can be obtained using least squares estimation procedures (Allen, 1997; Cohen &Cohen, 1983). The regression model from the dummy coding can be written as:

$$Y_{ij} = B_0 + \sum_{j=1}^{k-1} B_j D_{ij} + \varepsilon_{ij}.$$

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: The number of categories of the independent variable.

B_j : The regression coefficient associated with the jth group, and 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 assigned to subject i in the jth group.

ε_{ij} : The error associated with the ith subject in the jth group.

The general rule of coding states that all members of a given group are assigned identical numerical values. It follows that their predicted scores are also identical. The predicated score for each subject is equal to the mean of the group to which the subject belongs. In addition, the coefficient of multiple determination, R^2 , for the regression model with dummy variables can be interpreted in terms of the proportion of variance in the dependent variable that is accounted for by the categorical independent variable (Allen, 1997; Cohen & Cohen, 1983; Keppel & Zedeck, 1989; Myers & Well,2003).

3.3 Assumptions Underlying Structural Model

The regression model from dummy coding is based on the following assumptions (Myers & Well, 2003):

1) The errors are independently and normally distributed with mean 0 and variance σ_e^2.
2) All group means lie on a straight line.
3) The errors are not correlated with the independent variable.

3.4 Advantages and Disadvantages

The dummy coding is the preferred method when one wishes to compare several treatment group with a control group. In this case, the control group may serve as the reference group and the regression coefficients would then reflect the treatment-control mean differences (Myers & Well, 2003). However, this method does not test the differences between specific treatment means as well as the effect of a particular treatment defined as the deviation between the treatment mean and the grand mean (Cohen & Cohen, 1983).

3.5 An Example of Dummy-Coded Data

To illustrate the application of dummy coding in the regression analysis, consider a simple example in which the researcher was interested in the relationship between students exposure to different types of teaching methods and their performance on a standardized mathematics test. The students were divided into three groups, those who were taught by the discovery method, those who were taught by the observational method, and those who were taught by the traditional method. There were five students in each group.

In this situation, two dummy variables, D_1 and D_2, would be needed to classify the three groups. Table 1 displays the dummy coding of the independent variable teaching method. Note that subjects in the discovery group have been coded 1 for D_1 and 0 for D_2, those in the observational group have been coded 0 for D_1 and 1 for D_2, and those in the traditional group have been coded 0 for both D_1 and D_2. As such, the traditional group served as the reference group.

Table 1: Dummy coding for data of three groups

Group	Score	D_1	D_2
Discovery	90	1	0
	88	1	0
	91	1	0
	95	1	0
	93	1	0
Observational	78	0	1
	74	0	1
	71	0	1
	76	0	1
	70	0	1
Traditional	56	0	0
	59	0	0
	54	0	0
	55	0	0
	58	0	0

Table 2 summarizes results for the regression analysis of the mathematics scores on the teaching methods using dummy coding. As shown in Table 2, the intercept (B_0=56.40) represents the mean of the traditional group. This is because the intercept in the regression equation is equal to the expected value of the dependent variable whenever the values of the independent variables are equal to zero. In this case, the traditional group was coded 0 on all the dummy variables. Thus, the intercept represents the mean for this group. Similarly, the regression coefficient associated with D_1 (B_1=35.00) indicates that the mean of the discovery group is 35 points greater than that of the traditional group. This difference is statistically significant, $t(12) = 20.07$, $p = .000$. Also, the regression coefficient associated with D_2 (B_2 = 17.40) indicates that the mean of the observational group is 17.40 points greater than that of the traditional group. This difference is statistically significant, $t(12) = 9.98$, $p = .000$. The resulted estimated regression equation for the dummy coded data is:

$$\hat{Y} = 56.40 = 35.00D_1 + 17.40D_2.$$

Table 2: Regression of the mathematics scores on the teaching methods using dummy coding

Variable	*B*	*SE B*	β	*t*	*p*
D_1	35.00	1.74	1.14	20.07	.000
D_2	17.40	1.74	.57	9.98	.000
Constant	56.40				

However, no conclusion can be made regarding the difference between the mean of the discovery group to that of the observational group. If the researcher is interested in the difference between the means of these two groups, then one of the multiple comparisons tests should be conducted (pedhazur, 1997). In addition, the results of the regression analysis indicate that the proportion of variance in the mathematics scores accounted for by the teaching method (R^2 = .97) is statistically significant, $F(2,12) = 201.482$, $p = .000$.

It should be evident that, when dummy coding is used to code a categorical variable, the test of significance of a given regression coefficient is equivalent to a test of the difference between the mean of the group associated with the regression coefficient and the mean of the reference group. Also, it should be noted that the *F* ratio associated with the R^2 of the dependent variable with the dummy variables is equivalent to the overall *F* ratio for the test of the null hypothesis that the group population means are equal to each other. As such, in the present example, one may conclude that at least one group population mean is different from the others.

4. Effect Coding Method

4.1 Definition

In this method, the dummy variables take on the values 1, 0, and -1. Indeed, 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 0, but in the effect coding it is coded -1 (Cohen & Cohen, 1983; Myers & Well, 2003).

4.2 Structural model

When effect coding is used in the regression analysis, the overall results for the regression model (R^2 and *F*) are the same as in the dummy coding (Cohen & Cohen, 1983). However, the interpretations of the intercept and the regression coefficients are different. The regression model from the effect coding can be written as follows:

$$Y_{ij} = B_0 + \sum_{j=1}^{k-1} B_j E_{ij} + \varepsilon_{ij}.$$

Where

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

B_0: The intercept that represents the grand mean of the dependent variable for all groups.

k : The number of categories of the independent variable.

B_j: The regression coefficient associated with the jth group, and it represents the difference between the mean of the group coded 1 on the corresponding dummy variable and the grand mean of all groups. In other words, it represents the effect of being in the jth group. Hence, this method is named effect coding.

E_{ij}: The numerical value assigned to subject i in the jth group.

ε_{ij} : The error associated with the ith subject in the jth group.

4.3 Assumptions Underlying Structural Model

The assumption underlying the structural model from the effect coding are the same as in the dummy coding, which are:

1) The errors are independently and normally distributed with mean 0 and variance σ_e^2 .

2) All group means lie on a straight line.

3) The errors are not correlated with the independent variable.

4.4 Advantages and disadvantages

Effect coding is appropriate when each group is compared with the entire set of groups rather than with a reference group. In other words, effect coding is useful in testing the effect of a treatment defined as the deviation between the treatment mean and the grand mean. However, to determine which means differ significantly from each other, one of the methods for multiple comparisons of means has to be applied (Cohen & Cohen, 1983).

4.5 An Example of Effect-Coded Data

Table 3: Effect coding for data of three groups

Group	Score	E_1	E_2
Discovery	90	1	0
	88	1	0
	91	1	0
	95	1	0
	93	1	0
Observational	78	0	1
	74	0	1
	71	0	1
	76	0	1
	70	0	1
Traditional	56	-1	-1
	59	-1	-1
	54	-1	-1
	55	-1	-1
	58	-1	-1

Table 3 shows similar data to that shown in Table 1. However, this time the independent variable (i.e., teaching method) has been coded using the effect coding method. Similarly, two dummy variables, E_1 and E_2, are required to fully represent the information in the classification of the teaching method. Scores of the discovery group receive a 1 on E_1 and a 0 on E_2, scores of the observational group receive a 0 on E_1 and a 1 on E_2, and scores of the traditional group receive values of -1 on both E_1 and E_2.

Table 4 summarizes results for the regression analysis of the mathematics scores on the teaching methods using effects coding. The intercept ($B_{0=}$73.87) represents the grand mean of all groups. The regression coefficient associated with E_1($B_{1\,=}$ 17.53) indicates that the mean of the discovery group is 17.53 points greater than the grand mean of all groups. This difference is statistically significant, t(12) = 17.42, p=.000. Similarly, the regression coefficient associated with E_2 (B_2 = -0.07) indicates that the mean of the observational group is 0.07 points smaller than the grand mean of all groups. This difference is not statistically significant, t(12) = -0.07, p = .948. Also, the proportion of variance in the mathematics scores accounted for by the types of teaching methods (R^2 = .97) is statistically significant, F(2,12) = 201.48, p = .000; the same values as were obtained by the dummy coding method. The estimated regression equation from the effect-coded data is:

$$\widehat{Y} = 73.87 + 17.53E_1 - 0.07E_2.$$

Table 4: Regression of the mathematics scores on the teaching methods using effect coding

Variable	B	$SE\ B$	β	t	p
D_1	17.53	1.08	.99	17.42	.000
D_2	-0.07	1.08	-0.004	-0.07	.948
Constant	73.87				

Clearly, in effect coding, each regression coefficient reflects the effect of being in a particular group or treatment. Hence, testing the significance of the regression coefficient is equivalent to testing the significance of the treatment effect. Also, testing the significance of R is equivalent to testing the null hypothesis of no difference in the population between the groups' means.

5. Unequal Sample Sizes

Up to this point, the application of the coding methods in the regression analysis was limited to designs with equal sample sizes per group. Indeed, unequal sample sizes have no impact on the dummy coding method. However, for effect coding, unequal sample sizes produce a change in the interpretations of the intercept and regression coefficients. In such cases, the intercept represents the unweighted average of the group means. Similarly, the regression coefficients represent a comparison of the group mean coded 1 to the unweighted average of the group means (Cohen & Cohen, 1983; Myers & Well, 2003).

6. Analysis of Variance and Multiple Regression

The foregoing discussion has shown the relationship between analysis of variance and multiple regression. Indeed, the two statistical procedures are not entirely different. The analysis of variance examines whether the groups have different means, and provides an *F* ratio on the differences between the means. The multiple regression analysis examines whether the means are related to the groups, and yields an *F* ratio on the significance of R^2, which amounts to the same thing. However, analysis of variance may be considered as a special case of multiple regression. That is because multiple regression analysis can encompass both categorical and continuous variables, whereas analysis of variance is limited to categorical independent variables (Myers & Well, 2003; Pedhazur, 1997).

7. Summary

The purpose of this paper was to describe how categorical independent variables can be incorporated into regression analysis by virtue of two coding methods: dummy and effect coding. In general, for a given set of data, both methods yield identical R^2 and *F*. However, the two methods differ in the information provided by the regression equation. Table 5 contrasts dummy coding to effect coding with respect to the coding system, intercept, regression coefficients, uses, and effect of unequal sample sizes.

Table 5: Points of contrasts between dummy and effect coding

Points of contrasts	Dummy coding	Effect coding
Coding system	0 and 1	1, 0, and -1
Intercept	Mean of group coded all 0s ($\overline{Y}_0$)	Grand mean of all groups, j = 1, 2, …, k; ($\overline{Y}_{..}$)
Regression coefficient	$\overline{Y}_j - \overline{Y}_0$	$\overline{Y}_j - \overline{Y}_{..}$
Uses	Compare several experimental groups with a control group	Test treatment effect
Effect of unequal sample sizes	Unaffected by sample sizes	Intercept = unweighted average of the group means ($\overline{Y}_{un}$). Regression coefficient = $\overline{Y}_j - \overline{Y}_{un}$

References

Allen, M. P. (1997). *Understanding regression analysis* [Electronic version]. New York: Plenum Press.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation for the behavioral sciences (2nd ed.)* [Electronic version]. Hillsdale, NJ: Lawrence Erlbaum Associates.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches*. New York: W.H. Freeman and Company.

Myers, J. L., & Well, A.D. (2003). *Research design and statistical analysis* (2nd ed). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Grady, K. E., & Medoff, D. R. (1988). Categorical variables in multiple regression: Some cautions. *Multivariate Behavioral Research, 23,* 243-260. http://dx.doi.org/10.1207/s15327906mbr2302_7

Pedhazur, E. J. (1977). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace College.

Copyright Disclaimer