IOSR Journal of Mathematics (IOSRJM)
ISSN: 2278-5728 Volume 2, Issue 5 (Sep-Oct 2012), PP 01-07
www.iosrjournals.org

Use of Ordinal Dummy Variables in Regression Models

I.C.A. Oyeka[1], C.H. Nwankwo[2]
[1,2]*Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria.*

Abstract: *Many activities and phenomena on earth which are of interest to man and require to be studied are not quantitative in nature, they are rather qualitative. Sometimes their contributions, as independent variables, in a multiple regression, to variations in a specified quantitative dependent variable and to characterize them are of interest. Usually dummy variables with equal spacing are used in generating the design matrix despite its uninterpretable coefficients. Here a cumulatively coded design matrix is proposed and the coefficients are interpreted. This method is applied to an illustrative example, alongside two other possible methods, to demonstrate its applicability, and the proposed method showed a comparatively good performance with an additional advantage of interpretable coefficients which is very useful for practical purposes.*
Keywords:*cumulative, dummy, independent, ordinal, qualitative*

I. Introduction

Qualitative variables abound in many spheres of life and there are lots of interests in the activities which generate these variables. These interests include determining the relative contributions of the various levels of an independent variable in explaining the variations of a specified dependent variable. Interest may also be on characterising the nature of relationship between a quantitative dependent variable and a set of qualitative independent variables.

Usually the various levels of the independent variables of interest are assigned numerical codes with equal spacing. These codes may not however reflect the true pattern of relationships amongst the categories of the variables [1].

This paper developes a method of using ordinal dummy variables in multiple regression models, by a pattern of 1's and 0's, which avoids the restrictive requirement of equal spacing of levels of the independent variables , while achieving the objective of the model, or requiring any distributional form or requirement of homogeneity of variances.

II. The Proposed Method

In the ordinal dummy variable coding system each category or level of a parent independent variable in a regression model is represented ordinally by a pattern of 1's and 0's forming a dummy variable set. In order to avoid linear dependence among the dummy variables of a parent variable each parent variable is always represented by one dummy variable less than the number of its categories [2],[3]. Thus if a given parent variable Z has z categories or levels, the corresponding design matrix X will be represented ordinally by z-1 column vectors of ordinally coded dummy variables x_d of 1's and 0's (for d = 1, 2, …, z-1) The 1's and 0's in each x_d are cumulative if the values of the level of the parent variable it represented are arranged together.

Specifically, the pth level (p = 1, 2, … z) of the z levels of a parent variable Z will be represented by x_d ordinally coded column vector of 1's and 0's for d = 1, 2, … z-1 that is:

$$x_{id} = \begin{cases} 1, \text{ if } p > d \text{ for all } d = 1, 2 \ldots z-1, \text{ where } p\ (p=1, 2 \ldots z) \\ \text{Is the level of the parent variable Z in which the value} \\ y_i = y_{ip} \text{ of the dependent variable Y is reported} \\ 0, \text{ otherwise for } i = 1, 2 \ldots n \end{cases} \quad (1)$$

That means:

$$x_{id} = \begin{cases} 1, \text{ if } y_i \text{ is in level d of Z; } p > d;\ d = 1, 2 \ldots z-1;\ p = 1, 2 \ldots z \\ 0, \text{ otherwise} \end{cases} \quad (1a)$$

For example, suppose n observations are made on a parent variable Z having z levels with n_1 of the observations falling in level 1 of Z, n_2 observations falling in level 2 … and finally n_z falling in level z of Z, where $n = \sum_{i=1}^{z} n_i$. Then if, but without loss of generality the observations in each level of Z are arranged all together then the n x (z-1) design matrix X representing Z will consist of a set of z-1 cummulatively coded column vectors, x_d of 1's and 0's of the form

$$X = \begin{array}{c} \text{Level of } Z \\ 1 \\ \\ \\ \\ \\ \\ \\ \\ 2 \\ \\ \\ \\ \\ \\ \\ \\ 3 \\ \\ \\ \\ \\ z \\ \\ \end{array} \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1z-1} \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \quad \text{.....(2)}$$

Equation 2 is a prototype of ordinally coded design matrix X with z-1 cumulatively coded column vectors x_d of 1's and 0's representing the z levels of the parent variable Z. Note that the first n_1 elements of the first column x_1 of X representing the first level of Z are 0's while the remaining n-n_1 are all 1's. The first n_1 + n_2 elements of x_2 are 0's, while the remaining n-(n_1 + n_2) elements are all 1's and so on, until finally all the elements of x_{z-1} are all 0's except the last n_z elements which are all 1's.

Note that all the observations in the first level (level 1) of Z are all coded 0 in all the columns of the design matrix X while observations in the last level (level z) of Z are all coded 1's in X.

Note also that Z may be any set of parent independent variables such as A, B, C etc with levels a, b, c, etc respectively.

An ordinal dummy variable multiple regression model of y_i on the x_{ij}'s may be expressed as:

$$y_i = \beta_0 + \beta_{1;A}X_{i1;A} + \beta_{2;A}X_{i2;A} + \ldots\ldots + \beta_{a-1;A}X_{i,a-1;A} + \ldots\ldots\ldots\ldots + \beta_{c-1;C}X_{i,c-1;C} + e_i \quad (3)$$

Where β_j's are partial regression coefficients and e_i are error terms uncorrelated with x_{ij}'s, with E (e_i) = 0; A has 'a' levels, B has 'b' levels … C has 'c' levels, etc.

Note that the expected value of y_i is:

$$E(y_i) = \beta_0 + \beta_{1;A}X_{i1;A} + \beta_{2;A}X_{i2;A} + \ldots\ldots\ldots\ldots + \beta_{a-1;A}X_{i,a-1;A} + \ldots\ldots + \beta_{c-1;C}X_{i,c-1;C} + e_i \quad (4)$$

Equation 3 may alternatively be expressed in its matrix form as:

$$\underline{y} = X\underline{\beta} + \underline{e} \quad (5)$$

where $\underline{y}$ is an nx1 column vector of outcome values; X is an nxr cumulatively coded design matrix of 1's and 0's; $\underline{\beta}$ is an rx1 column vector of regression coefficients and $\underline{e}$ is an nx1 column vector of error terms uncorrelated with X with E($\underline{e}$) = $\underline{0}$ where r is the rank of the design matrix X.

Use of the method of least squares with either equation (3) or (5) yields an unbiased estimator of _ as:

$$\underline{\beta} = \underline{b} = (X'X)^{-1}X'\underline{y} \quad (6)$$

Where $(X'X)^{-1}$ is the matrix inverse of $(X'X)$, the resulting predicted regression model is:

$$\underline{\hat{y}} = X\underline{b} \quad (7)$$

The following analysis of variance (ANOVA) table (Table 1) enables the testing of the adequacy of equations 3 or 5 using the F test.

TABLE 1: Analysis of Variance (ANOVA) Table for Equation 5

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Sum of Squares (MS)	F. Ratio
Regression	$SSR = \underline{b}'x'\underline{y} - n\bar{y}^2$	$r-1$	$MSR = \frac{SSR}{r-1}$	$F = \frac{MSR}{MSE}$
Error	$SSE = \underline{y}'\underline{y} - \underline{b}'x'\underline{y}$	$n-r$	$MSE = \frac{SSE}{n-r}$	
Total	$SST = \underline{y}'\underline{y} - n\bar{y}^2$	$n-1$		

The null hypothesis to be tested for the adequacy of the regression model (Equation 3 or 5) is:

$$H_0 : \underline{\beta} = \underline{0} \; versus \; H_1 : \underline{\beta} \neq \underline{0} \quad (8)$$

H_0 is tested using the test statistic $F = \frac{MSR}{MSE}$

which has an F distribution with r-1 and n-r degrees of freedom. H_0 is rejected at the α level of significance if:

$$F \geq F_{(1-\alpha;\ r-1,\ n-r)} \quad (9)$$

Otherwise H_0 is accepted where $F_{(1-\square\ ;\ ₹1,\ n-r)}$ is the critical value of the F distribution with r-1 and n-r degrees of freedom for a specified α level.

If H_0 is rejected indicating that not all β_j's are equal to zero, then some other hypotheses concerning β_j's may be tested.

Note that β_k; is interpreted in ordinal dummy variable regression model as the amount by which the dependent variable y on the average changes for every unit increase in x_k compared with x_{k-1} or one unit decrease in x_k relative to x_{k+1} when all other independent variables in the model are held constant[4]. That is β_k measures the amount by which on the average the dependent variable y increases or decreases for every unit change in x_K compared with a corresponding unit change in either x_{k-1} or x_{k+1} respectively when all other independent variables in the model are held constant.

Research interest may be in comparing the differential effects of any two ordinal dummy variables of a parent independent variable on the dependent variable. For example one may be interested in testing the null hypothesis:

$$H_0 : \beta_{l;A} = \beta_{j;A} \ versus \ H_1 : \beta_{l;A} > \beta_{j;A} \quad (10)$$

Where the β_d's are estimated from Eqn (3.7) as b_d's for l = 1, 2,… , a-1; j = 1, 2, … ,a-1; $l \neq$ j.

The null hypothesis of Equation (10) may be tested using the test statistic:

$$t = \frac{b_{l;A} - b_{j;A}}{se\left(b_{l;A} - b_{j;A}\right)} = \frac{b_{l;A} - b_{j;A}}{\sqrt{\left(\underline{C}\left(X'X\right)^{-1}\underline{C}^{1}\right)MSE}} \quad (11)$$

Where $\underline{C}$ is an r row vector of the form (0, 0… 1, 0…-1, 0…0)

Where 1 and -1 correspond to the positions of $\beta_{l;A}$ *and* $\beta_{j;A}$ respectively in the rx1 column vector $\underline{b}$ and all other elements of $\underline{C}$ are 0; $\left(X'X\right)^{-1}$ is the matrix inverse of $\left(X'X\right)$. H_0 is rejected at the level of significance $\square$ if: $t \geq t_{(1-\alpha;\ n-r)}$,

(12) otherwise H_0 is accepted, where $t_{(1-\alpha;\ n-r)}$ is the critical value of the t distribution with n-r degrees of freedom for a specified α level.

In general several other hypotheses may be tested. For example one may be interested in comparing the effects of the ith level of factor A, say and the jth level of factor C, say, or of some combinations of some levels of several factors. Thus interest may be in testing:

$$H_0 : \beta_{l;A} = \beta_{j;C} \ versus \ H_1 : \beta_{l;A} \neq \beta_{j;C} \quad (13)$$

Using the test statistic

$$t = \frac{b_{l;A} - b_{j;C}}{se\left(b_{l;A} - b_{j;C}\right)} = \frac{\underline{C}\underline{b}}{\sqrt{\left(\underline{C}\left(X'X\right)^{-1}\underline{C}^{1}\right)MSE}} \quad (14)$$

Where $\underline{C}$ is a row vector as in Equation 11, except that 1 and -1 now occurs at the positions corresponding to the ith level of factor A and jth level of factor C in $\underline{b}$. H_0 is rejected as in Equation 12.

Further interest may also be in estimating the total or overall effect of a given parent independent variable through the effects of its representative ordinal dummy variables on the dependent variable. To do this it should be noted that any parent variable is completely determined by its set of representative ordinal dummy variables.

III. Illustrative example

A Clinician collected data on age, gender, duration of infection and packed cell volume (PCV) of a random sample of 80 HIV-positive patients shown in table 2 below. Interest is in determining the effects of age, gender and duration of infection of the PCV levels of HIV-positive patients.

TABLE 2: Data on Random Sample of HIV-Positive Patients

S/N	Age (Year)	Sex	Duration of Infection (Year)	PCV Level
1	28	M	.5	32
2	27	F	1.0	27
3	39	M	6.0	30
4	40	F	5.0	32
5	26	M	5.0	33
6	31	M	.5	36
7	71	M	2.0	24
8	58	F	2.0	29
9	62	F	1.1	24
10	63	F	2.6	27
11	27	F	3.0	32
12	61	F	7.0	27
13	61	F	2.7	35
14	32	F	1.8	36
15	32	M	1.8	46
16	26	F	1.7	27
17	36	F	3.0	28
18	35	M	2.4	30
19	45	M	3.8	35
20	33	M	2.3	38
21	38	F	2.5	28
22	39	M	.4	30
23	45	M	2.1	30
24	32	F	.1	28
25	40	M	.4	32
26	32	M	.3	42
27	57	M	.2	36
28	29	F	.5	31
29	27	F	.7	24
30	46	F	.3	34
31	45	M	.6	27
32	32	F	3.0	35
33	32	M	2.5	34
34	28	F	.3	17
35	38	M	3.5	40
36	30	F	1.7	30
37	28	F	4.4	30
38	28	M	2.3	37
39	45	F	2.8	26
40	30	M	1.6	35
41	27	M	3.1	34
42	30	F	.3	34
43	25	F	4.1	28
44	25	F	.5	27
45	20	F	.1	31
46	65	M	.4	30
47	52	M	4.1	27
48	26	F	1.1	28
49	24	F	1.9	34
50	60	F	2.6	33
51	33	M	1.8	36
52	31	F	.1	29
53	31	M	.9	49
54	30	M	2.6	40
55	33	F	2.6	35
56	42	F	2.1	34
57	25	F	2.6	37
58	31	F	1.8	29
59	23	F	1.4	33
60	32	F	.3	24
61	28	F	.4	38
62	38	F	.1	29

63	37	M	1.8	33
64	31	F	.3	32
65	36	M	.1	40
66	38	M	.8	31
67	35	F	.4	25
68	43	M	.5	29
69	42	M	1.7	39
70	36	F	.8	31
71	32	F	2.0	24
72	29	F	4.0	28
73	25	F	1.6	36
74	47	M	2.2	37
75	27	F	.6	14
76	40	M	3.3	41
77	30	F	2.4	31
78	38	M	2.3	32
79	28	F	2.8	33
80	35	F	5.4	28

To use ordinal dummy variables to represent the parent independent variables age, gender, and duration of infection we may group age into four classes, namely 20-29years(1), 30-39years(2), 40-49years(3), and 50years or more(4) and duration of infection into four groups, namely, less than 1year(1), 1- 2years(2), 2 – 3years(3) and above 3 years(4). Using these classifications in equation 1 with the data of table 2 we obtain the ordinal variable representation of table 2 as table 3 below.

Table 3

32.00	1.00	.00	.00	.00	1.00	.00	.00	.00
27.00	1.00	.00	.00	.00	.00	.00	.00	.00
30.00	1.00	1.00	.00	.00	1.00	1.00	1.00	1.00
32.00	1.00	1.00	1.00	.00	.00	1.00	1.00	1.00
33.00	1.00	.00	.00	.00	1.00	1.00	1.00	1.00
36.00	1.00	1.00	.00	.00	1.00	.00	.00	.00
24.00	1.00	1.00	1.00	1.00	1.00	1.00	.00	.00
29.00	1.00	1.00	1.00	1.00	.00	1.00	.00	.00
24.00	1.00	1.00	1.00	1.00	.00	1.00	.00	.00
27.00	1.00	1.00	1.00	1.00	.00	1.00	1.00	.00
32.00	1.00	.00	.00	.00	.00	1.00	1.00	.00
27.00	1.00	1.00	1.00	1.00	.00	1.00	1.00	1.00
35.00	1.00	1.00	1.00	1.00	.00	1.00	1.00	.00
36.00	1.00	1.00	.00	.00	.00	1.00	.00	.00
46.00	1.00	1.00	.00	.00	1.00	1.00	.00	.00
27.00	1.00	.00	.00	.00	.00	1.00	.00	.00
28.00	1.00	1.00	.00	.00	.00	1.00	1.00	.00
30.00	1.00	1.00	.00	.00	1.00	1.00	1.00	.00
35.00	1.00	1.00	1.00	.00	1.00	1.00	1.00	1.00
38.00	1.00	1.00	.00	.00	1.00	1.00	1.00	.00
28.00	1.00	1.00	.00	.00	.00	1.00	1.00	.00
30.00	1.00	1.00	.00	.00	1.00	.00	.00	.00
30.00	1.00	1.00	1.00	.00	1.00	1.00	1.00	.00
28.00	1.00	1.00	.00	.00	.00	.00	.00	.00
32.00	1.00	1.00	1.00	.00	1.00	.00	.00	.00
42.00	1.00	1.00	.00	.00	1.00	.00	.00	.00
36.00	1.00	1.00	1.00	1.00	1.00	.00	.00	.00
41.00	1.00	.00	.00	.00	.00	.00	.00	.00
24.00	1.00	.00	.00	.00	.00	.00	.00	.00
34.00	1.00	1.00	1.00	.00	.00	.00	.00	.00
27.00	1.00	1.00	1.00	.00	1.00	.00	.00	.00
35.00	1.00	1.00	.00	.00	.00	1.00	1.00	1.00
34.00	1.00	1.00	.00	.00	1.00	1.00	1.00	.00
17.00	1.00	.00	.00	.00	.00	.00	.00	.00
40.00	1.00	1.00	.00	.00	1.00	1.00	1.00	1.00

30.00	1.00	1.00	.00	.00	.00	1.00	.00	.00
30.00	1.00	.00	.00	.00	.00	1.00	1.00	1.00
37.00	1.00	.00	.00	.00	1.00	1.00	1.00	.00
26.00	1.00	1.00	1.00	.00	.00	1.00	1.00	.00
35.00	1.00	1.00	.00	.00	1.00	1.00	.00	.00
34.00	1.00	.00	.00	.00	1.00	1.00	1.00	.00
34.00	1.00	1.00	.00	.00	.00	.00	.00	.00
28.00	1.00	.00	.00	.00	.00	1.00	1.00	1.00
27.00	1.00	.00	.00	.00	.00	.00	.00	.00
31.00	1.00	.00	.00	.00	.00	.00	.00	.00
30.00	1.00	1.00	1.00	1.00	1.00	.00	.00	.00
27.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
28.00	1.00	.00	.00	.00	.00	1.00	.00	.00
34.00	1.00	.00	.00	.00	.00	1.00	.00	.00
33.00	1.00	1.00	1.00	1.00	.00	1.00	1.00	.00
36.00	1.00	1.00	.00	.00	1.00	1.00	.00	.00
29.00	1.00	1.00	.00	.00	.00	.00	.00	.00
41.00	1.00	1.00	.00	.00	1.00	.00	.00	.00
40.00	1.00	1.00	.00	.00	1.00	1.00	1.00	.00
35.00	1.00	1.00	.00	.00	.00	1.00	1.00	.00
34.00	1.00	1.00	1.00	.00	.00	1.00	1.00	.00
37.00	1.00	.00	.00	.00	.00	1.00	1.00	.00
29.00	1.00	1.00	.00	.00	.00	1.00	.00	.00
33.00	1.00	.00	.00	.00	.00	1.00	.00	.00
24.00	1.00	1.00	.00	.00	.00	.00	.00	.00
38.00	1.00	.00	.00	.00	.00	.00	.00	.00
29.00	1.00	1.00	.00	.00	.00	.00	.00	.00
33.00	1.00	1.00	.00	.00	1.00	1.00	.00	.00
32.00	1.00	1.00	.00	.00	.00	.00	.00	.00
40.00	1.00	1.00	.00	.00	1.00	.00	.00	.00
31.00	1.00	1.00	.00	.00	1.00	.00	.00	.00
25.00	1.00	1.00	.00	.00	.00	.00	.00	.00
29.00	1.00	1.00	1.00	.00	1.00	.00	.00	.00
39.00	1.00	1.00	1.00	.00	1.00	1.00	.00	.00
31.00	1.00	1.00	.00	.00	.00	.00	.00	.00
24.00	1.00	1.00	.00	.00	.00	1.00	1.00	.00
28.00	1.00	.00	.00	.00	.00	1.00	1.00	1.00
36.00	1.00	.00	.00	.00	.00	1.00	.00	.00
37.00	1.00	1.00	1.00	.00	1.00	1.00	1.00	.00
14.00	1.00	.00	.00	.00	.00	.00	.00	.00
41.00	1.00	1.00	1.00	.00	1.00	1.00	1.00	1.00
31.00	1.00	1.00	.00	.00	.00	1.00	1.00	.00
32.00	1.00	1.00	.00	.00	1.00	1.00	1.00	.00
33.00	1.00	.00	.00	.00	.00	1.00	1.00	.00
28.00	1.00	1.00	.00	.00	.00	1.00	1.00	1.00

Two known regression procedures were applied on the data, these are the use of the real values of age, duration of infection and sex(dummy: 1 for male, 0 for female), the second is the use of the normal dummy variable coding with equal spacing of levels in both age and duration of infection using the intervals in the example. The proposed ordinal cummulative coding method in this paper is also used. Table 4 below show summary of the results obtained by the three different methods.

TABLE 4

	Real Values	Normal Coding	Ordinal Coding
R^2	.238	.279	.277
F-value	7.926(p-value 0.000)	3.985(p-value 0.001)	3.939(p-value 0.001)

The real values show a smaller R^2 value compared with the two coding methods. On the other hand, the normal coding method with equal spacing show a slightly higher R^2 value (.279) over that for the ordinal cumulative coding method (R^2 = .277).The three of them however show significant F-values with small p-values, desirable properties.

From the fore goings, the two coding methods outperformed the use of raw values. The normal coding method, though with a marginal edge in the R^2 value, by .002, may not be preferred to the ordinal cumulative coding. This is because the coefficients of the normal coding method do not have clear interpretations of the regression coefficients due to the restriction of equal spacing of levels of the independent variables. On the other hand, the cumulative coding regression coefficients proposed here can be interpreted. For instance, the β coefficients using the proposed cumulative coding method for age 30-39 is -0.673, this is interpreted as the decrease, on the average, in PCV level, due to the age interval 30-39 relative to the age interval 20-29 or increase in PCV level on the average due to the age interval 30-39 relative to the age interval 40-49, when all other independent variables are held constant.

IV. Conclusion

The performance of the ordinal coding method is therefore of relatively high quality, not only by the R^2 values, the interpretable regression coefficients make it more suitable for practical purposes. Its robust nature, as in the case of other coding techniques, makes it outstanding.

References

[1] I.C.A. OYEKA (1993), Estimating effects in ordinal dummy variable regression, **"**STATISTICA, anno LIII, n. 2" pp. 262-268.
[2] R. P. BOYLE (1970), Path analysis and ordinal data, "American Journal of Sociology", 47, 1970, 461-480.
[3] J. NETER, W. WASSERMAN, M. H. KUTNER (1983), Applied linear regression models (Richard D. Irwin Inc, Illinois).
[4] M. LYONS (1971), Techniques for using ordinal measures in regression and path analysis, in Herbert Costner (ed.)(Sociological Methods , Josey Bass Publishers, San Francisco).