American Journal of Theoretical and Applied Statistics
2014; 3(2): 44-48
Published online February 20, 2014 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20140302.13

Direct and indirect effects in dummy variable regression

Oyeka I. C. A., Nwankwo Chike H.*

Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria

Email address:
chikeezeoke@yahoo.com (N. Chike H.)

To cite this article:
Oyeka I. C. A., Nwankwo Chike H.. Direct and Indirect Effects in Dummy Variable Regression. *American Journal of Theoretical and Applied Statistics.* Vol. 3, No. 2, 2014, pp. 44-48. doi: 10.11648/j.ajtas.20140302.13

Abstract: This paper proposes and develops the use of the non-cummulative dummy variables of 1's and 0's to represent levels of parent independent variables in dummy variable multiple regression models. The regression coefficients obtained using the proposed methods are easier to interprete and clearly understand than the use of the cummulatively coded ordinal dummy variables of 1's and 0's that could be used for the same purpose. The proposed method also enables the simultaneous estimation of the total, absolute or overall effect of a parent independent variable as well as its direct effect through its representative dummies and its indirect effect on a given independent variable through the mediation of other parent independent variables in the model was demonstrated. The use of these procedures was illustrated with an example.

Keywords: Dummy Variables, Total Effect, Direct Effect, Indirect Effect, Parent Variables, Mediation Model

1. Introduction

Cumulatively coded ordinal dummy variables of 1's and 0's have been used to represent independent variables in a regression model [1]. Estimates of the partial regression effects of these dummy variables on a given independent variable are provided. Interpretation of these regression coefficients is often difficult and not easily understood, this is because the regression coefficient of an ordinal dummy variable representing a given level of a parent independent variable is interpreted as the effect on the dependent variable per unit change in the level of the parent independent variable represented by that dummy variable in comparism or relative to a unit increase in the level of the parent independent variable represented by an immediately preceding ordinal dummy variable or per unit decrease in the level of the parent independent variable represented by an immediately succeeding ordinal dummy variable [2]. This interpretation is rather cumbersome.

2. The Proposed Method

An easier to interprete and understand method is to perhaps use the more regular non-cumulative dummy variables of 1's and 0's to represent levels of parent independent variables in a regression model. It may also be of interest to use this method to estimate the total, absolute or overall effect of a parent independent variable as well as its direct effect through its representative dummies and its indirect effect on a given independent variable through the mediation of other parent independent variables in the model as discussed below.

Here, the dependent variable may or may not be quantitative. The dependent variable or the so-called parent independent variables may also each be either a quantitative or qualitative variable. But for the present purpose, each of these parent independent variables that is not already categorical is to be partitioned into a number of mutually exclusive categories, classes or levels.

Now suppose y_i is the score, value, observation on ith subject on a criterion or dependent variable Y, for i = 1, 2, …, n. Suppose further that the effects of characteristics A, B, C, … etc, with levels a, b, c, … of each subject with respect to the dependent variable are of interest. To use these parent independent variables in a dummy variable regression model, we would represent each of them with one dummy variable of 1's and 0's less than the number of its levels or categories. This is to ensure that the resulting design matrix is of full column rank and hence non-singular ([2], [3], and [4]). Thus if the parent independent variable A has 'a' levels coded 1, 2, .., a, then these 'a' levels are represented by a-1 dummy variables of 1's and 0's in the regression model, that is by the dummy variables $x_{i1;A}, x_{i2;A}, \ldots, x_{ia-1;A}$ for the ith subject, i = 1, 2, …, n. Other parent independent variables are similarly represented. Thus factor B with 'b' levels is represented by b–1 dummy variables of 1's and 0's and factor C with c levels is represented by c–1 dummy variables of 1's and 0's,

and so on.

Now a dummy variable multiple regression model expressing the dependence of scores or observations y_i drawn from the criterion variable Y on the parent independent variables A, B and C represented by dummy variables $x_{i1;A}, x_{i2;A}, ..., x_{ia-1;A}, ..., x_{ic-1;C}$, is expressed as

$$y_i = \beta_0 + \beta_{1;A}x_{i1;A} + \beta_{2;A}x_{i2;A} + ... + \beta_{a-1;A}x_{ia-1;A} + \beta_{1;B}x_{i1;B} + ... + \beta_{c-1;C}x_{ic-1;C} + ... + e_i \quad (1)$$

where β_j's are partial regression coefficients and e_i's are error terms uncorrelated with x_{ij}'s with $E(e_i) = 0$, for i = 1, 2, …, n. The partial regression coefficient β_j of the dummy variable x_j representing the jth level of a given parent independent variable is interpreted as the change in the dependent variable per unit change in the jth level of that parent independent variable relative to or in comparism with its other levels holding all other parent independent variables in the model at constant levels [2].

The expected value of y_i of equation (1) is

$$E(y_i) = \beta_0 + \beta_{1;A}x_{i1;A} + \beta_{2;A}x_{i2;A} + ... + \beta_{a-1;A}x_{ia-1;A} + \beta_{1;B}x_{i1;B} + ... + \beta_{c-1;C}x_{ic-1;C} + ... \quad (2)$$

Use of the usual least squares method with equation (1) gives the fitted or predicted dummy variable multiple regression model as

$$\hat{y}_i = b_0 + b_{1;A}x_{i1;A} + b_{2;A}x_{i2;A} + ... + b_{a-1;A}x_{ia-1;A} + b_{1;B}x_{i1;B} + ... + b_{c-1;C}x_{ic-1;C} \quad (3)$$

for i = 1, 2, … n

Note that if in equation (1) or (3) we set some selected dummy variables equal to 1 and some equal to 0, several other models and estimates of the parameters of these models describing the dependence of the criterion variable on various combinations of the levels of the different parent independent variables used in the model are obtained, thereby further highlighting the tremendous versatility and usefulness of dummy variable regression models in statistical modelling.

Use of the usual F test enables one assess the adequacy of a hypothesised model in correctly describing the true pattern of relationships between the dependent variable and the set of parent independent variables used in the model.

If the model fits, that is, if the model is adequate, leading to a rejection of the null hypothesis, in which case not all the regression coefficients are zero, then one may proceed to test other hypotheses and also estimate additional parameters of interest, including absolute direct and indirect effects of parent independent variables on the criterion variable.

Now to estimate the so called direct effect [5] which is actually the partial regression coefficient or effect of a given parent independent variable Z say with z levels on a criterion or dependent variable Y, we treat the dummy variables representing the parent independent variable Z as intermediate variables between Z and Y in a regression model. Then following the method of path analysis we obtain the required direct effect of Z on Y as a weighted sum of the partial regression coefficients of the dummy variables used as regressor on the dependent variable. For example, we obtain the direct effect $\beta_{dir;A}$ of the parent independent variable A on each of its representative dummy variables $x_{ij;A}$, i = 1, 2, … n; j = 1, 2, ….., a - 1

Specifically, to determine the partial regression coefficient or the so-called direct effect of the parent independent variable A on the dependent variable Y, we take the partial derivative of the expected value of y_i of equation 1, that is of equation 2 with respect to A, obtaining

$$\beta_{dir;A} = \frac{dE(y_i)}{dA} = \beta_{1;A}\frac{dE(x_{i1;A})}{dA} + \beta_{2;A}\frac{dE(x_{i2;A})}{dA} + ... + \beta_{a-1;A}\frac{dE(x_{ia-1;A})}{dA} + \sum_s \beta_{S;Z}\frac{dE(x_{is;Z})}{dA}$$

s = 1, 2, ….. for all parent independent variables Z different from A.

or

$$\beta_{dir;A} = \sum_{j=1}^{a-1} \beta_{j;A}\frac{dE(x_{ij;A})}{dA} \quad (4)$$

Since $\sum_s \beta_{S;Z}\frac{dE(x_{is;Z})}{dA} = 0$

because $\frac{dE(x_{is;Z})}{dA} = 0$

Now the weight $\alpha_{j;A} = \frac{dE(x_{ij;A})}{dA}$, to be applied to $\beta_{j;A}$, the partial regression effect of the dummy variable $x_{ij;A}$ representing the jth level of the parent impendent variable A on the dependent variable Y is obtained by fitting a simple regression model of A regressing on $x_{ij;A}$

for j = 1, 2, … , a-1 using assigned numerical codes.

Thus for the dummy variable $x_{ij;A}$ representing the jth level of the parent independent variable A we obtain the weight α_j by fitting the simple regression line

$$x_{ij;A} = \alpha_{0;A} + \alpha_{j;A}A \tag{5}$$

for i = 1, 2, … n; j = 1, 2, …, a -1 and $E(e_i) = 0$

Now taking the partial derivative of the expected value of $x_{ij;A}$ of equation 5 with respect to A, we obtain

$$\frac{dE(x_{ij;A})}{dA} = \alpha_{j;A} \quad \text{for j = 1, 2, …, a -1} \tag{6}$$

The direct effect which is actually the partial regression effect of the parent independent variable A on the dependent variable Y is now obtained by using equation 6 in equation 4 as

$$\beta_{dir;A} = \sum_{j=1}^{a-1} \alpha_{j;A}\beta_{j;A} \tag{7}$$

whose sample estimate is obtained as a weighted sum of the sample estimates of the partial regression coefficients $\hat{\beta}_{j;A} = b_{j;A}$ as

$$b_{dir;A} = \sum_{j=1}^{a-1} \alpha_{j;A} b_{j;A} \tag{8}$$

An advantage of using dummy variables to represent independent variables in a multiple regression model is that it enables separate estimation of the partial effect of each level or category of a parent independent variable on a dependent variable which clearly provides additional information. It also enables the simultaneous estimation of not only the direct effects as we have already seen, but also the total or absolute effect and the indirect effect of a parent independent variable on a dependent variable through the mediation of other parent independent variables in the regression model.

The indirect effect of a given parent independent variable on a dependent variable is the difference between its total or absolute effect and its direct effect through its representative dummy variables. The total or absolute effect itself is the simple regression coefficient or regression effect of the parent independent variable using directly its assigned numerical codes on the dependent variable.

Thus the indirect effect $\beta_{ind;A}$ of the parent independent variable A on a dependent variable Y through the mediation of other parent independent variables in the model is estimated as its total or absolute effect $b_{t;A}$ less its direct effect $b_{dir;A}$. That is

$$b_{ind;A} = b_{t;A} - b_{dir;A} \tag{9}$$

where $b_{t;A}$ is the sample estimate of the simple regression coefficient or effect of the parent independent variable A using its numerical codes on the dependent variable Y.

The total, direct and indirect effects of other parent independent variables in the model (Equation 1) are similarly estimated.

3. Illustrative Example

A researcher is interested in estimating the effects of maternal age (A), mother's body weight (W) and her parity (P) has on birth weight (B = y) of her most recent live birth.

She collected a random sample of 25 newly delivered mothers as shown in table 1 below

Table 1: *Child birth weight and some demographic factors of a random sample of 25 mothers*

S/N	Mother's Age (A)	Parity (P)	Mother's Body Weight (W)	Baby's Weight (B;y)
1	31	2	75	3.5
2	26	1	63	3.9
3	30	3	55	3.5
4	28	9	90	2.8
5	24	4	90	3.0
6	36	1	76	3.8
7	20	1	59	2.7
8	34	1	73	3.7
9	25	1	62	2.8
10	25	2	83	2.9
11	30	5	75	3.0
12	24	2	68	3.2
13	21	1	72	3.4
14	26	3	78	2.6
15	22	1	72	2.7
16	28	1	68	3.1
17	39	7	65	3.6
18	28	3	61	3.2
19	28	5	70	2.8
20	18	2	72	2.8
21	19	1	73	2.9
22	21	1	78	3.2
23	28	4	75	3.1
24	28	3	68	3.0
25	17	1	65	2.8

To use dummy variable multiple regression methods with data of table 1 above we first partition age of mother (A) into three groups or levels as (1) < 25years (2) 25 – 29 years (3) ≥ 30 years[(1), (2) and (3) being parent independent variables representing the various age ranges]; Parity into three classes or groups as (1) 0 or 1, (2) 2 – 3 (3) 4 or more, [(1), (2) and (3) being the parent independent variables for indicated parity]. Mother's body weight is grouped into two classes (1) ≤ 70kg, (2) > 70kg. [Similarly, (1) and (2) are parent independent variables for the two weight intervals]. Hence maternal age (A) and parity (P) each with 3 levels will each be represented by 2 dummy variables of 1's and 0's while Mother's body weight with 2

levels will be represented with only 1 dummy variable of 1's and 0's in the dummy variable regression model.

The resulting design matrix is

Table 2: *Design Matrix X of Dummy variables for the data on table 1*

S/No	$X_{i1;A}$	$X_{i2;A}$	$X_{i1;P}$	$X_{i2;P}$	$X_{i1;W}$	Baby's Weight
1	0	0	0	1	0	3.5
2	0	1	1	0	1	3.9
3	0	0	0	1	1	3.5
4	0	1	0	0	0	2.8
5	1	0	0	0	0	3.0
6	0	0	1	0	0	3.8
7	1	0	1	0	1	2.7
8	0	0	1	0	0	3.7
9	0	1	1	0	1	2.8
10	0	1	0	1	0	2.9
11	0	0	0	0	0	3.0
12	1	0	0	1	1	3.2
13	1	0	1	0	0	3.4
14	0	1	0	1	0	2.6
15	1	0	1	0	0	2.7
16	0	1	1	0	1	3.1
17	0	0	0	0	1	3.6
18	0	1	0	1	1	3.2
19	0	1	0	0	1	2.8
20	1	0	0	1	0	2.8
21	1	0	1	0	0	2.9
22	1	0	1	0	0	3.2
23	0	1	0	0	0	3.1
24	0	1	0	1	1	3.0
25	1	0	1	0	1	2.8

Hence the fitted dummy variable multiple regression model expressing the dependence of child birth on maternal age, body weight and parity represented by dummy variables is

$$\hat{y}_i = 3.387 - 0.622x_{1;A} - 0.512x_{2;A} - 0.512x_{2;A} + 0.243x_{1;P} + 0.078x_{2;P} + 0.069x_{1;W} \quad (10)$$

Now, for the direct effects of the parent independent variables on Y (ie baby weight), we obtain the regression coefficients (α_j's) each of $x_{i1;A}$ on A, $x_{i2;A}$ on A, $x_{i1;P}$ on P, $x_{i2;P}$ on P with the following results

$x_{i1;A}$ vs A yields $x_{i1;A} = 2.192 - 0.070A$

$$\therefore \alpha_{1;A} = -0.07 \quad (11)$$

$x_{i2;A}$ vs A yields $x_{i2;A} = 0.122 - 0.011A$

$$\therefore \alpha_{2;A} = -0.011 \quad (12)$$

$x_{i1;P}$ vs P yields $x_{i1;P} = 0.872 - 0.166P$

$$\therefore \alpha_{1;P} = -0.166 \quad (13)$$

$x_{i2;P}$ vs P yields $x_{i2;P} = 0.340 - 0.008P$

$$\therefore \alpha_{2;P} = -0.008 \quad (14)$$

Also $x_{i1;W}$ vs W yields $x_{i1;W} = 3.710 - 0.046W$

$$\therefore \alpha_{1;W} = -0.046 \quad (15)$$

Thus, the direct effect of mothers age (A) on the baby's weight (y) from equation (8) using equation (10) and equations (11) and (12), is

$$b_{dir;A} = (-0.622 \times -0.07) + (-0.5 \times -0.011) = 0.049 \quad (16)$$

Similarly, the direct effect of parity (P) on baby's weight (y) using equations (8), (13) and (14) is

$$b_{dir;P} = (0.243 \times -0.166) + (0.078 \times -0.008) = -0.041 \quad (17)$$

And finally, the direct effect of mother body weight (W) on baby's weight (y) using equation (8), (10) and (15) is

$$b_{dir;W} = 0.069 \times -0.041 = -0.003 \quad (18)$$

For the total effects, the dependent variable, baby's weight (y), is regressed on each of the parent independent variables 1, 2 and 3 for the three levels of mother's age (A), another 1, 2 and 3 parent independent variables for the three levels of parity (P) and two levels of mother's body weight (W) represented by 1 and 2.

These regressions produced the following equations

$$Y = 2.453 + 0.333A \text{ (for mother's Age)} \quad (19)$$

$$Y = 3.244 - 0.069P \text{ (for Parity)} \quad (20)$$

and

$$Y = 3.191 - 0.045W \text{ (for Mother's Weight)} \quad (21)$$

The coefficients for A, P and W respectively from equations (19), (20) and (21) are the total effects of age, parity and mother's weight.

The indirect effects for mother's age, parity and mother's weight on baby's weight are obtained using equation (9), thus indirect effect for mother's age is

$$b_{ind;A} = b_{t;A} - b_{dir;A} = 0.333 - 0.049 = 0.284 \quad (22)$$

For parity, it is

$$b_{ind;p} = b_{t;p} - b_{dir;p} = -0.069 - -0.041 = -0.028 \quad (23)$$

and lastly, the indirect effect of mother's weight on baby's weight is

$$b_{ind;w} = -0.045 - -0.003 = -0.042 \quad (24)$$

4. Conclusion

We have presented a method for the estimation of total or absolute effects and the direct effect of parent independent variables on a dependent variable through the effects of their set of representative dummy variables of 1's and 0's as well as indirect effects of these parent independent variables through the mediation of other parent independent variable in a dummy variable regression model.

It is shown that an advantage of using these dummy variables of 1's and 0's is the ease of understanding and interpretation of the resulting estimated regression coefficients in comparison with the regression coefficients obtained when cummulatively coded ordinal dummy variables of 1's and 0's are used in such models.

References

[1] Oyeka I. C. A. and Nwankwo, C. H. (2012): "Use of Ordinal Dummy Variables in Regression Models". IOSR Journal of Mathematics, Vol. 2, Issue 5 (Sep – Oct 2012), pp 10 – 07.

[2] Boyle, R. P. (1970): "Path Analysis and Ordinal Data". American Journal of Sociology, 47, 1970, 461 – 480.

[3] Oyeka I. C. A. (1993): "Estimating Effects in Ordinal Dummy Variable Regression". STATISTICA, anno L.111, n.2 pp 262 – 268.

[4] Neter, J., Wasserman, W. and Kutner, M. H. (1983): "Applied Linear Regression Models". Richard D. Irwin Inc, Illinois.

[5] Wright, S. (1960): "Path Coefficients and Path Regression: Alternative to Complementary Concepts". Biometrics, Volume 16, pp 189 – 202.