Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.4, No.7, 2014 

# Use of Dummy Variable Regression Techniques in Estimating Joint and Marginal Values of an Outcome

Nwankwo Chike H. (Corresponding Author)
Department of Statistics, Nnamdi Azikiwe University, P M B 5025, Awka , Anambra State, Nigeria, E-mail of corresponding author: chikeezeoke@yahoo.com

Oyeka, I.C.A.
Department of Statistics , Nnamdi Azikiwe University, P M B 5025, Awka
Anambra State, Nigeria, E-mail: cyprianoyeka@yahoo.com

**ABSTRACT**

This paper proposes a statistical method for estimating the values, including the joint and marginal values of an outcome variable, using dummy variable multiple regression techniques. Estimates are also provided for the direct and indirect effects of a parent independent variable on a dependent variable through its representative dummy variables in the presence of other independent variables in the model. The proposed method is illustrated with some data.

**Keywords:** Dummy Variables, Total Effect, Direct Effect, Indirect Effect, Parent Variables, Mediation Model, Joint Value, Marginal Value.

## INTRODUCTION

The use of dummy variables in a regression model enables the researcher estimate the separate effects of the levels of a parent independent variable on an outcome or dependent variable. These effects may then be used to estimate the values of the dependent variables at various levels of specified independent variables of interest.

In this paper we propose to use these techniques to provide estimates of the joint and marginal values of an outcome or dependent variable at various levels of given parent independent variable. Estimates of the so called "direct" and "indirect" effects of a given parent independent variable on a dependent variable are also provided.

### The Proposed Method

To use the dummy variable approach in a multiple regression model we first partition each of the parent independent variables if not already categorical, into a set of mutually exclusive categories, levels or subgroups and then use dummy variables of 1's and 0's to represent these categories in a regression model. In such a regression model each parent independent variable is represented by one dummy variable of 1's and 0's less than the number of its categories. This is to avoid linear dependence among the columns of the deign matrix X of the regression model and hence ensure that X is of full column rank; r (Boyle 1970; Neter and Wasserman 1974; Oyeka 1992; Hardy 1993).

Thus interest may be in determining the relationship between a dependent variable Y and a set of parent independent variables or factors A B C ….. with levels a b c … respectively. However for greater clarity, simplicity and ease of interpretation, but without loss of generality, we will here present the proposed method for only three parent independent variables A, B, and C, assumed to have levels a, b, and c respectively.

Now let $y_i$ be the response or score of the ith randomly selected subject on the outcome or dependent variable Y for i = 1, 2 . . . n. To use the dummy variable Y for i = 1, 2, . . . n. To use the dummy variable multiple regression method to estimate the effects of the parent independent variables or factors A, B and C on the dependent variable Y we let

 

$x_{ij;A}$ = 1, if the observation on the ith subject is in level j of factor A, j = 1, 2 … a-1
0, otherwise

$x_{ij;B}$ = 1, if the observation on the ith subject is in level j of factor B, j = 1, 2 … b-1
0, otherwise

$x_{ij;C}$ = 1, if the observation on the ith subject is in level j of factor C, j = 1, 2 … c-1
0, otherwise

for i = 1, 2, … n – 1

Any level of a parent variable that is not represented by a dummy variable is termed the "omitted" or "excluded level" for that parent variable. Thus levels a, b, and c are termed the omitted levels of variables A, B and C respectively.

The dependence of $y_i$ on these dummy variables may be expressed by the dummy variable multiple regression model

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

for i = 1, 2, … n.

where the $\beta_j$ 's are regression coefficients and $e_i$'s are error terms uncorrelated with the dummy variables $x_{ij}$ 's with $E(e_i) = 0$, for i = 1, 2, …, n.

Note that the expected value of Equation 2 is

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

Equation 2 may alternatively be expressed in its matrix form as

$$\underline{y} = X\underline{\beta} + \underline{e} \quad . \quad . \quad . \quad (4)$$

Where $\underline{y}$ is an nx1 column vector of response scores or observations, X is an n x r design matrix of 1's and 0's of full column rank r; $\underline{\beta}$ is an r x 1 column vector of regression coefficients; and $\underline{e}$ is an n x 1 column vector of error terms uncorrelated with X with E($\underline{e}$) = $\underline{O}$ where r is the rank of X or the number of parameters (regression coefficients) in the model.

 

The method of least squares is used with either Equation 2 or 4 to obtain unbiased estimates of the regression co-efficient as

$$\underline{\hat{\beta}} = \underline{b} = (X'X)^{-1} X' \underline{y} \quad . \quad . \quad . \quad (5)$$

Where $(X'X)^{-1}$ is the matrix inverse of $(X'X)$

This yields the predicted dummy variable multiple regression model for $\underline{y}$ as

$$\underline{\hat{y}} = X\underline{b} \quad . \quad . \quad . \quad (6)$$

which may alternatively be expressed as

$$\hat{y}_i = b_0 + b_{1;A}x_{i1;A} + b_{2;A}x_{i2;A} + \ldots\ldots + b_{a-1;A}x_{ia-1;A} + b_{1;B}x_{i1;B} + b_{2;B}x_{i2;B} + \ldots\ldots + b_{b-1;B}x_{ib-1;B} + b_{1;C}x_{i1;C} + b_{2;C}x_{i2;C} + \ldots\ldots + b_{c-1;C}x_{ic-1;C} \quad \ldots \quad (7)$$

For i = 1, 2 … n

To estimate the value of the dependent variable Y for given levels of a parent independent variable compared with its omitted or excluded level, we set the dummy variables representing the given levels of that parent independent variable equal to 1 and all dummy variables representing all other parent independent variables in the model equal to 0 in Equation 7. Thus the estimated value of the dependent variable Y at the a-1 levels of the parent independent variable A is obtained by setting.

$$x_{i1;A} = x_{i2;A} + \ldots\ldots + x_{ia-1;A} = 1 \; and \; x_{i1;B} = x_{i2;B} = \ldots\ldots = x_{ib-1;B} = x_{i1;C} = x_{i2;C} = \ldots\ldots x_{ic-1;C} = 0$$

in Equation 7, yielding

$$\hat{y}_i = b_0 + b_{1;A} + b_{2;A} + \ldots\ldots + b_{a-1;A} \quad . \quad . \quad . \quad (8)$$

Note that $b_{j;A}$ is interpreted as the estimated effect on the dependent variable y of the jth level of factor A compared with the other levels of A when other independent variables in the model are held at constant levels for j = 1, 2 . . . a-1. Other estimated regression coefficients in Equation 7 are similarly interpreted.

Now the predicted joint value of Y at the hth level of A, jth level of B and lth level of C is obtained from Equation 7 by setting

$x_{ih;A} = x_{ij;B} = x_{il;C}$ = 1 and all other $x_{iv}'s = 0$

for all $v \neq h, j$ and $l \; (h = 1,2...a-1; \; j = 1,2...b-1; \; l = 1,2.....c-1)$

yielding

$$\hat{y}_i = b_0 + b_{h;A} + b_{j;B} + b_{l;C} \quad . \quad . \quad . \quad (8)$$

For h = 1, 2 . . . a-1; j = 1, 2, . . . b – 1; and l = 1, 2 . . . c-1

The estimated joint value of Y at the omitted level of A; The jth level of B and the lth level of C is obtained by setting.

$x_{ih;A} = 0$ for all h = 1, 2 . . . a-1; $x_{ij;B} = x_{il;C} = 1$ and

$x_{iv;B} = x_{iv;C} = 0$ for $v \neq h, j, l \; (h = 1,2...a-1; \; j = 1,2...b-1; \; l = 1,2.....c-1)$

 

in Equation 7 yielding

$$\hat{y}_i = b_0 + b_{j;B} + b_{l;C} \quad . \quad . \quad . \quad (9)$$

for j = 1, 2 . . . b-1; l = 1, 2, . . . c – 1

Other estimated joint values of Y are similarly obtained

The estimated marginal value of $y_i$ at the hth level of factor A and lth level of factor C for all levels of factor B is obtained by setting

$$x_{ih;A} = x_{il;C} = 1; x_{iv;A} = x_{iv;C} = 0 \text{ for } v \neq h, l \quad (h = 1,2...a-1;\ l = 1,2.....c-1)$$

and $x_{ij;B} = 1$ for all j = 1, 2 . . . b-1; in Equation 7 yielding

$$\hat{y}_i = b_0 + b_{h;A} + b_{l;C} + \sum_{j=1}^{b-1} b_{j;B} \quad . \quad . \quad . \quad (10)$$

for h = 1, 2 . . . a-1; l = 1, 2, . . . c – 1

The marginal value of $y_i$ at the hth level of factor A and all levels of factors B and C is estimated by setting

$$x_{ih;A} = 1;\quad x_{iv;A} = 0;\quad v \neq h \quad (h = 1,2,...,a-1;) \text{ and } x_{iv;B} = 1; x_{iv;C} = 1;$$

For all $v = j, l \quad (j = 1,2,...,b-1;\ l = 1,2,...,c-1)$ in Eqn 7 giving

$$\hat{y}_i = b_0 + b_{h;A} + \sum_{j=1}^{b-1} b_{j;B} + \sum_{l=1}^{c-1} b_{l;C} \quad . \quad . \quad . \quad (11)$$

h = 1, 2 . . . a-1;

The marginal value of $y_i$ at the omitted value of A and lth level of C for all levels of B is estimated by setting

$x_{ih;A} = 0;$ for all h = 1, 2 . . . a-1; $b_{l;C} = 1; b_{v;C} = 0;\ v \neq l \quad (l = 1,2,...,c-1;)$ and $b_{j;B} = 1;$ for all j = 1,2, . . ., b-1 in Eqn 7 giving

$$\hat{y}_i = b_0 + b_{l;C} + \sum_{j=1}^{b-1} b_{j;B} \quad . \quad . \quad . \quad (12)$$

The marginal value of $y_i$ at the omitted levels of factors A and C is estimated by setting

$$x_{iv;A} = x_{iv;C} = 0 \text{ for } v = h, l \quad (h = 1,2,...,a-1; l = 1,2,...,c-1); \text{ and}$$

$x_{ij;B} = 1$ for all j = 1, 2 . . . b-1 in Eqn 7 yielding

$$\hat{y}_i = b_0 + \sum_{j=1}^{b-1} b_{j;B} \quad . \quad . \quad . \quad (13)$$

All other marginal values of $y_i$ are similarly estimated from Equation 7.

Although it is highly illuminating to examine separately the effects of various levels of a parent independent variable through its representative dummy variables and the resulting values of the dependent variable itself, it

 

may also be of research interest to determine the absolute or total, direct and indirect effects of such a parent independent variable on the dependent variable.

The total or absolute effect of a parent independent variable on a dependent variable is the regression coefficient obtained by fitting a simple linear regression of the dependent variable on the parent independent variable. The direct effect of a parent independent variable is the weighted sum of the partial regression coefficients or effects of its representative dummy variables on the dependent variable in the presence of other independent variables in the model. The corresponding indirect effect is a measure of the effect of the parent independent variable on the dependent variable through the mediation of other independent variables in the model and is measured as the difference between its total and direct effects (Wright 1960).

Now to obtain the direct effect $B_A$ of a given parent independent variable A on a dependent variable Y, we treat the dummies representing the parent independent variable A as intermediate variables between A and y.

Then following the method of path analysis (Wright 1960, Lyons 1971) we obtain the direct effect $B_A$ as a weighted sum of the partial regression co-efficients $\beta_{h;A}$, h = 1, 2 . . . a-1; from Equation 3.

Specifically the weight $\alpha_{h;A}$ to be applied to $\beta_{h;A}$ is obtained by fitting a regression line of $x_{ih;A}$ on the parent variable A. Thus for the hth dummy variable $x_{ih;A}$ representing the parent variable A, we fit the regression line.

$$x_{ih;A} = \beta_0 + \alpha_{h;A}A + e_{ih;A} \quad . \quad . \quad . \quad (14)$$

Now taking the partial derivative of the expected value of Equation 14 with respect to A we obtain

$$\frac{dE(x_{ih;A})}{dA} = \alpha_{h;A} \quad . \quad . \quad . \quad (15)$$

for h = 1, 2 . . . a-1

Now the partial regression effect or the so called direct effect of the parent independent variable A through the effects of its representative dummy variables on the dependent variable y in the presence of other parent independent variables in the model is obtained by taking the partial derivative of the expected value of $y_i$ (Equation 3) with respect to A. That is

$$\beta_A = \frac{\partial E(y_i)}{\partial A} \quad . \quad . \quad . \quad (16)$$

Now

$$\frac{dE(y_i)}{dx_{ih;A}} = \beta_{h;A}$$ (see Equation 3); h = 1, 2 . . . a-1

Hence using Equation 3 we have that

$$\beta_A = \beta_{1;A} \cdot \frac{dE(x_{1;A})}{dA} + \beta_{2;A} \cdot \frac{dE(x_{2;A})}{dA} + \ldots + \beta_{a-1;A} \cdot \frac{dE(x_{ia-1;A})}{dA} + \beta_{1;B} \frac{dE(x_{i1;B})}{dA} + \beta_{2;B} \frac{dE(x_{i2;B})}{dA} + \ldots$$

$$+ \beta_{c-1;C} \frac{dE(x_{ic-1;C})}{dA}$$

Now

$$\frac{dE(x_{ih;A})}{dA} = \alpha_{h;A}$$ (see Equation 16) for h = 1, 2 . . . a-1 and

 

$$\frac{dE(x_{ij;B})}{dA} = \frac{dE(x_{il;C})}{dA} = 0 \qquad \text{for j = 1, 2 . . . b-1; l = 1, 2 . . . c -1}$$

for all parent variables B and C different from A.

Hence the direct effect of the parent independent variable A on the dependent variable y through its representative dummy variables is given as

$$\beta_A = \alpha_{1;A} \cdot \beta_{1;A} + \alpha_{2;A} \cdot \beta_{2;A} + ... + \alpha_{a-1;A} \cdot \beta_{a-1;A} + 0$$

or

$$\beta_A = \sum_{h=1}^{a-1} \alpha_{h;A} \cdot \beta_{h;A} \qquad . \quad . \quad . \qquad (17)$$

Whose sample estimate is

$$\beta_A = \sum_{h=1}^{a-1} \alpha_{h;A} \cdot b_{h;A} \qquad . \quad . \quad . \qquad (18)$$

The direct effects of B and C may be similarly estimated.

The difference between the total regression effect $B_A$ that is the simple regression coefficient of y on the parent independent variable A and $\beta_A$, the direct effect of A on y through its representative dummy variables in the presence of other parent independent variables in the model provides a measure of the so called indirect effect of the parent independent variable A on y through the mediation of other parent independent variables in the model. That is

$$\beta_{A/B,C} = B_A - \beta_A \qquad . \quad . \quad . \qquad (19)$$

estimated as

$$\hat{\beta}_{A/B,C} = \hat{B}_A - b_A \qquad . \quad . \quad . \qquad (19)$$

**Illustrative Example**

Table 1 presents data on the Packed Cell Volume (PCV), Age Duration of infection and Gender of a random sample of 80 HIV positive patients from a certain community.

 

**Table 1:** Data on a Random Sample of HIV positive Patients

| S/No | PCV | Age (years) | Duration (years) | Sex |
|---|---|---|---|---|
| 1 | 32 | 28 | 0.5 | M |
| 2 | 27 | 27 | 1.3 | F |
| 3 | 30 | 39 | 6.3 | M |
| 4 | 32 | 40 | 5.3 | F |
| 5 | 33 | 26 | 5.3 | M |
| 6 | 36 | 31 | 0.5 | M |
| 7 | 24 | 71 | 2.0 | M |
| 8 | 29 | 58 | 2.0 | F |
| 9 | 24 | 62 | 1.0 | F |
| 10 | 27 | 63 | 2.6 | F |
| 11 | 32 | 27 | 3.0 | F |
| 12 | 27 | 61 | 7.0 | F |
| 13 | 35 | 61 | 2.7 | F |
| 14 | 36 | 32 | 1.8 | F |
| 15 | 46 | 32 | 1.8 | M |
| 16 | 27 | 26 | 1.7 | F |
| 17 | 28 | 36 | 3.0 | F |
| 18 | 30 | 35 | 2.4 | M |
| 19 | 35 | 45 | 3.8 | M |
| 20 | 38 | 33 | 2.2 | M |
| 21 | 28 | 39 | 2.5 | F |
| 22 | 30 | 39 | 0.4 | M |
| 23 | 30 | 45 | 2.0 | M |
| 24 | 28 | 32 | 0.1 | F |
| 25 | 32 | 40 | 0.3 | M |
| 26 | 42 | 32 | 0.6 | M |
| 27 | 36 | 57 | 0.4 | M |
| 28 | 31 | 29 | 0.2 | F |
| 29 | 24 | 27 | 0.7 | F |
| 30 | 34 | 46 | 0.3 | F |
| 31 | 27 | 45 | 0.6 | M |
| 32 | 35 | 32 | 5.0 | F |
| 33 | 34 | 32 | 2.5 | M |
| 34 | 17 | 28 | 0.2 | F |
| 35 | 40 | 38 | 3.5 | M |
| 36 | 30 | 30 | 1.7 | F |
| 37 | 38 | 28 | 4.4 | F |
| 38 | 37 | 28 | 2.2 | M |
| 39 | 26 | 45 | 2.8 | F |
| 40 | 35 | 30 | 1.6 | M |
| 41 | 34 | 27 | 3.1 | M |
| 42 | 34 | 30 | 0.2 | F |
| 43 | 28 | 25 | 4.1 | F |
| 44 | 27 | 25 | 0.5 | F |
| 45 | 31 | 20 | 0.1 | F |
| 46 | 30 | 65 | 0.4 | M |
| 47 | 27 | 52 | 4.1 | M |
| 48 | 28 | 36 | 1.1 | F |
| 49 | 34 | 24 | 1.9 | F |
| 50 | 33 | 60 | 2.6 | F |
| 51 | 36 | 33 | 1.9 | M |
| 52 | 29 | 31 | 0.1 | F |
| 53 | 41 | 31 | 0.9 | M |
| 54 | 40 | 30 | 2.6 | M |

 

| | | | | |
|---|---|---|---|---|
| 55 | 35 | 36 | 2.6 | F |
| 56 | 34 | 42 | 2.1 | F |
| 57 | 37 | 25 | 2.6 | F |
| 58 | 29 | 31 | 1.9 | F |
| 59 | 33 | 23 | 1.4 | F |
| 60 | 24 | 32 | 0.3 | F |
| 61 | 38 | 28 | 0.4 | F |
| 62 | 29 | 38 | 0.1 | F |
| 63 | 33 | 37 | 1.8 | M |
| 64 | 32 | 37 | 0.2 | F |
| 65 | 40 | 36 | 0.1 | M |
| 66 | 31 | 38 | 0.9 | M |
| 67 | 25 | 35 | 0.4 | F |
| 68 | 29 | 43 | 0.5 | M |
| 69 | 39 | 42 | 1.7 | M |
| 70 | 31 | 36 | 0.9 | F |
| 71 | 24 | 32 | 2.0 | F |
| 72 | 28 | 29 | 4.0 | F |
| 73 | 36 | 25 | 1.6 | F |
| 74 | 37 | 47 | 2.2 | M |
| 75 | 14 | 27 | 0.6 | F |
| 76 | 41 | 40 | 3.3 | M |
| 77 | 31 | 30 | 2.4 | F |
| 78 | 32 | 38 | 2.3 | M |
| 79 | 33 | 28 | 2.8 | F |
| 80 | 28 | 35 | 5.4 | F |

Interest is fitting a dummy variable multiple regression model with Packed Cell Volume (PCV) as the dependent variable and age, duration of infection and sex of HIV positive patients as parent independent variables and using the results obtained to illustrate the proposed method. To do this we here classify age in years into five groups or levels namely (1) less than 30 years (2) 30 – 34 years (3) 35 – 39 years (4) 40 – 49 years and (5) 50 years or more. Duration of infection in years is also classified into four groups namely (1) less than 1 year; (2) 1 year or more but less than 2 years; (3) 2 years or more but less than 3 years and (4) 3 or more years. Sex is classified into two levels namely (1) male and (0) female.

This means that in the dummy variable representation of parent variables and consistent with Equation 1 age (A) which has five levels would be represented by four dummy variables namely $x_{i1;A}$, $x_{i2;A}$, $x_{i3;A}$, and $x_{i4;A}$, for age levels (1) less than 30 years, (2) 30 – 34 years; (3) 35 – 39 years and (4) 40 – 49 years respectively. Duration of infection (B) with four levels will be represented by three dummy variables namely (1) $x_{i1;B}$, (2) $x_{i2;B}$, and (3) $x_{i3;B}$, for durations of infection levels (1) less than 1 year, (2) 1 – 2 years (3) 2 – 3 years respectively. Sex (C) which has two levels will be represented by 1 dummy variable, $x_{i1;C}$ for the male sex. Thus age level 50 years or more, duration level 3 years or more and the female gender are treated as the 'omitted' levels for age (A), duration of infection (B) and sex (C) respectively.

Using these specifications in Equation 1 for the data of Table 1 we obtain an 80 x 9 design matrix X of 1's and 0's using this design matrix and y to represent the dependent variable PCV in a dummy variable multiple regression model we obtain the fitted regression equation

$$\hat{y}_i = 27.3 + 2.14x_{i1;A}, + 4.52\,x_{i2;A} + 1.33x_{i3;A} + 2.65x_{i4;A} - 1.20x_{i1;B} + 0.81x_{i2;B} + 0.69x_{i3;B} + 4.96x_{i1;C} \quad \ldots(21)$$

Note from Equation 21 that with an estimated partial regression coefficient of $b_{1;A} = 2.14$ for age group less than 30 years for example indicates that for a given duration of infection and sex a randomly selected HIV

 

positive patient aged less than 30 years is on the average likely to have a PCV level of 2.14 higher than the PCV level of a randomly selected patient in other age groups.

Setting $x_{i2;A} = x_{i1;C} = 1$ and all other $x_{ij}'s = 0$ in Equation 21 yields

$$\hat{y}_i = 27.3 + 4.52 + 4.96 = 36.78$$

Interpreted, this means that a randomly selected male HIV patient age 30 – 34 years with a duration of infection of three or more years is estimated to have a PCV level of 36.78. A female HIV positive patient of the same characteristics is likely to have a PCV level of $27.3 + 4.52 = 31.82$

Now to estimate the joint PCV level of for example a male patient aged less than 30 years who has had the infection for less than one year, following Equation 8, we set

$x_{i1;A} = x_{i1;B} = x_{i1;C} = 1$ and all other $x_{ij}'s = 0$ in Equation 21 giving

$$\hat{y}_i = 27.3 + 2.14 - 1.20 + 4.96 = 33.2$$

The corresponding joint value for female patients is estimated by also now setting $x_{i1;C} = 0$ in Equation 21 obtaining

$\hat{y}_i = 27.3 + 2.14 - 1.20 = 28.24$. Thus a male HIV positive patient aged less than 30 years is likely to have a PCV level of 33.2 – 28.24 = 4.96 higher than the PCV level of his female counterpart.

The estimated joint PCV level of a male patient aged 50 years or more with a duration of infection of less than one year is obtained following Equation 9 by setting.

$x_{i1;A} = x_{i2;A} = x_{i3;A} = x_{i4;A} = x_{i2;B} = x_{i3;B} = 0$ and $x_{i1;A}; x_{i1;C} = 1$ in Equation 21 yielding

$$\hat{y}_i = 27.3 - 1.20 + 4.96 = 31.06$$

The corresponding value for his female counterpart is $\hat{y}_i = 27.3 - 1.20 + 4.96 = 26.1$. Thus a female HIV Positive patient aged 50 years or over with a duration of infection of less than one year is likely to have a PCV level of 4.96 less than that of her male counterpart

Similarly the point PCV level of a male patient aged fifty years or more with a duration of infection of 3 years or more is estimated using Equations 9 and 21 as $\hat{y}_i = 27.3 + 4.96 = 32.26$. The corresponding value for the female counterpart is 27.3, resulting in a male-female PCV difference of 4.96

The estimated marginal PCV level for a randomly selected male HIV positive patient aged less than 30 years no matter the duration of infection is obtained following Equation 11 by setting $x_{i1;A} = x_{i1;C} = 1;\ x_{i2;A} = x_{i3;A} = x_{i4;A} = 0$

$x_{i1;B} = x_{i2;B} = x_{i3;B} = 1$ in Equation 21 yielding

$$\hat{y}_i = 27.3 + 2.14 - 1.20 + 0.81 + 0.69 + 4.96 = 34.7$$

The corresponding PCV value for the female counterpart is obtained from Equation 21 by further setting $x_{i1;C} = 0$ giving

 

$\hat{y}_i = 27.3 + 2.14 - 1.20 + 0.81 + 0.69 + 4.96 = 34.7$, resulting in a male-female difference of estimated marginal PCV level of 4.96.

The estimated marginal PCV level for a randomly selected male HIV positive patient irrespective of duration of infection is obtained following Equation 12 by setting.

$x_{ih;A} = 0$ for all h = 1, 2, 3, 4 $x_{i1;C} = 1$ and $x_{ij;B} = 1$ for j = 1, 2, 3 in Equation 21 giving

$$\hat{y}_i = 27.3 - 1.20 + 0.81 + 0.69 + 4.96 = 32.56$$

The corresponding value for the female patients is obtained by further setting in Eqn 21 giving

$$\hat{y}_i = 27.3 - 1.20 + 0.81 + 0.69 = 27.6$$

Other marginal PCV values are similarly estimated and the results are shown in Table 2 below.

**Table 2:** Estimated Joint and Marginal PCV values of a sample of HIV positive patients by Age, Duration of Infection and Sex

| | **Duration of Infection** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Age** | **< 1 year (-1.20)** | | **1- 2 years (0.81)** | | **2 – 3 years (0.69)** | | **3 years or more (27.3)** | | **Total** | |
| | **Male (4.96)** | **Female** | **Male (4.96)** | **Female** | **Male (4.96)** | **Female** | **Male (4.96)** | **Female** | **Male (4.96) (marginal** | **Female for age)** |
| < 30 years (2.14) | 33.2 | 28.24 | 35.21 | 30.25 | 35.09 | 30.13 | 34.4 | 29.44 | 34.7 | 29.74 |
| 30 – 34 yrs (4.52) | 35.58 | 30.62 | 37.59 | 32.63 | 37.47 | 32.51 | 36.78 | 31.82 | 37.08 | 32.42 |
| 35 – 39 yrs (1.33) | 32.39 | 27.43 | 34.40 | 29.44 | 34.28 | 29.32 | 33.59 | 28.63 | 33.89 | 28.93 |
| 40 – 49 yrs (2.65) | 33.71 | 28.75 | 35.72 | 30.76 | 35.60 | 30.64 | 34.91 | 29.95 | 35.21 | 30.25 |
| 50 yrs + (27.3) | 31.06 | 26.10 | 33.07 | 28.11 | 32.95 | 27.99 | 32.26 | 27.30 | 32.56 | 27.6 |
| Total (Marginal for Duration) | 41.70 | 36.74 | 43.71 | 38.75 | 43.59 | 38.63 | 42.90 | 37.94 | 32.26 | 27.30 |

Now to estimate the direct and indirect effects of age, A; duration of illness B and sex C on PCV levels y, we as indicated above first regress each $x_{ih;A}(h = 1,2,3,4)$ on A; each $x_{ij;B}(j = 1,2,3)$ on B; $x_{ij;C}$ on C and $y_i$ on A, B and C. These will yield the estimated regression coefficient:

$\hat{\alpha}_{1;A} = -0.229;\ \hat{\alpha}_{2;A} = -0.076;\ \hat{\alpha}_{3;A} = 0.045;\ \hat{\alpha}_{4;A} = 0.084;$

$\hat{\alpha}_{1;B} = -0.311;\ \hat{\alpha}_{2;B} = -0.042;\ \hat{\alpha}_{3;B} = 0.106;\ \hat{\alpha}_{1;C} = 1.00;\ \hat{\beta}_A = -0.140;\ \hat{\beta}_B = 0.434;$ and

$\hat{\beta}_C = 4.84;$

Using these results in Equation (18) together with the estimated partial regression coefficient in Equation 21 we obtain the estimated direct effect of A on PCV levels y through its representative dummy variables as

 

$$b_A = (2.14)(-0.229)+(4.52)(-0.076)+(1.33)(0.045)+(2.65)(0.084) = -0.552$$

Furthermore the indirect effect of A on y through the mediation of other independent variables in the model is estimated from Equation 19 as

$$\widehat{B}_{A/B,C} = -0.140-(-0.552) = 0.412$$

These results show that in the absence of other independent variables in the model age has a subtractive effect on PCV level with its absolute effect estimated as

$\hat{\beta}_A = -0.140$. This means in this case that for every one year increase in the age of HIV patient the PCV level is on the average expected to reduce by 0.140 units. An estimated direct effect of age of $b_A = -0.5521$ on PCV level through its four representative dummy variables holding duration of infection and sex at constant levels, that is at specified levels of these two independent variables means that on the average, patients' PCV level is reduced by 0.552 units for every one year increase in age. However the indirect effect of age on PCV level through the mediation of duration of infection and sex at $\widehat{\beta}_{A/B,C} = 0.412$ is rather additive.

The direct and indirect effects of duration of infection B and sex C on PCV levels are similarly calculated and the results are presented in Table 4.

**Table 3:** Estimates of Absolute, Direct and Indirect Effect of Selected factors on PCV levels.

| | Estimated Effects | | |
|---|---|---|---|
| **Factor (Parent Variable)** | **Absolute (Total)** | **Direct** | **Indirect** |
| Age (years) | -0.140 | -0.552 | 0.412 |
| Duration of Infection | 0.434 | 0.412 | 0.022 |
| Sex | 4.84 | 4.96 | -0.12 |

## CONCLUSION

In this paper, statistical methods are presented for estimating the values as well as the joint and marginal values of an outcome variable using dummy variable multiple regression techniques. Estimates are also provided for the absolute, direct, and indirect effects of a parent independent variable on a dependent variable through the effects of its representative dummy variables in the presence of other independent variables in the model.

The illustrative example used shows that the proposed method can highly illuminate and clearly discriminate between the effects of various levels of a parent independent variable on a dependent variable in the presence of other independent variables in a regression model.

## REFERENCES

1) Boyle, R.P. (1970); "Path Analysis and Ordinal data". American Journal of Sociology, Volume 47, pp 461 – 480.
2) Hardy, M.A. (1993). "Regresion with Dummy Variables". Quantitative Applications in the social Sciences. A Sage University Paper. SAGE Publications, Newbury Park London.
3) Lyons, M. (1971); "Techniques for using Ordinal Measures in Regression and Path Analysis", in Hubert Costner (ed), Sociological Methods, Josey Bass Publishers, San Francesco.
4) Neter, J, Wasserman, W and Kutner, M. H. (1983); Applied Linear Regression Models. Richard D. Irwin Inc. Homewood Illinois.
5) Oyeka I.C.A. (1993): "Estimating Effects of Ordinal Dummy Variable Regression". STATISTICA, anno L. 111, n. 2, 1993 pp 261 – 8.
6) Wright, S. (1960); "Path Coefficients and Path Regression: Alternative or Complementary Concept". Biometrics, Volume 16, pp 189 – 202.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

## CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

**Prospective authors of journals can find the submission instruction on the following page:** http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

## MORE RESOURCES

Book publication information: http://www.iiste.org/book/

### IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar