European Journal of Statistics and Probability
Vol.2, No.1, pp.23-27, March 2014
Published by European Centre for Research Training and Development UK (www.ea-journals.org)

INCORPORATING DUMMY VARIABLES IN REGRESSION MODEL TO DETERMINE THE AVERAGE INTERNALLY GENERATED REVENUE AND WAGE BILLS OF THE SIX GEOPOLITICAL ZONES IN NIGERIA

Alabi Oluwapelumi

ABSTRACT: *This paper compare the year 2012 Internally Generated Revenue (IGR) and Wage bills of the six (6) Geopolitical zones in Nigeria. The Internally Generated Revenue and wage bills from the thirty six (36) states were categorized to six (6) Geopolitical zones (Southwest, Southsouth, Southwest, north central, northeast and northwest) and proxy with dummy variables. The average Internally Generated Revenue and Wage bills of each Geopolitical zone were derived from the expectation of the regression model and the estimated coefficient of its slope indicate which of the averages is statistically significant different. It also appraised which of the six geopolitical*

KEYWORDS*:* Dummy Variables, Regression Model, Wage Bills, Geopolitical Zones in Nigeria

INTRODUCTION

In statistics and econometrics, particularly in regression analysis, a dummy variable is an artificial variable created to represent an attribute with two or more distinct categories. It takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Dummy variables are used as devices to sort data into mutually exclusive categories. For example, in econometrics time series analysis, dummy variables may be used to indicate the occurrence or non occurrence of event. A dummy variable can thus be thought of as a truth value represented as a numerical value 0 or 1 (as is sometimes done in computer programming).

In a regression model, a dummy variable with a value of 0 will cause its coefficient to disappear from the equation. Conversely, the value of 1 causes the coefficient to function as a supplemental intercept, because of the identity property of multiplication by 1. This type of specification in a linear regression model is useful to define subsets of observations that have different intercepts and/or slopes without the creation of separate models. In logistic regression models, encoding all of the independent variables as dummy variables allows easy interpretation and calculation of the odds ratios, and increases the stability and significance of the coefficients. Synonyms for dummy variables are design variables [Hosmer and Lemeshow, 1989], Boolean indicators, and proxies [Kennedy, 1981]. Related concepts are binning [Tukey, 1977] or ranking, because belonging to a bin or rank could be formulated into a dummy variable. Bins or rankscan also function as sets and dummy variables can represent non-probabilistic set membership. Set theory is usually explained in texts on computer science or symbolic logic. See [Arbib, et. al., 1981] or [MacLane, 1986]. Dummy variables based on set membership can help when there are too few observations, and thus, degrees of freedom, to have a dummy variable for every category or some categories are too rare to be statistically significant.

Analysis of variance models are used to assess the statistical significance of the relationship between a quantitative regressand and qualitative or dummy regressors. They are often used to compare the differences in the mean values of two or more groups or categories, and are therefore more general than the t test which can be used to compare the means of two groups or categories only. Influence the regressand and clearly should be included among the explanatory variables, or the regressors. Since such variables usually indicate the presence or absence of a "quality" or an attribute, they are essentially nominal scale variables. One way we could "quantify" such attributes is by constructing artificial variables that take on values of 1 or 0, 1 indicating the presence (or possession) of that attribute and 0 indicating the absence of that attribute. Such variables are thus essentially a device to classify data into mutually exclusive categories.Dummy variables can be incorporated in regression models just as easily as quantitative variables. As a matter of fact, a regression model may contain regressors that are all exclusively dummy, or qualitative, in nature. Such models are called Analysis of Variance models**.**For an applied treatment, see John Fox, Applied Regression Analysis, Linear Models, and Related Methods, Sage Publications, 1997.There are various statistical techniques to compare two or more mean values, which generally go by the name of analysis of variance.

MATERIALS AND METHOD

The dataset for this paper is derived from daily trust investigation and the statistical bulletin of the national bureau of statistics. The data is the record of year 2012 Internally Generated Revenue and wage bills of the 36 states in Nigeria, the states were categorized to the six (6) geopolitical zone.
1.southwest zone (6 states),
2.south south zone (6 states),
3.southeast zone (5 states),
4.north central (7 states) including FCT

5.northwest zone (7 states),

6.northeast zone (6 states) and

Analysis of variance model

$$Y_i = \beta_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 d_4 + \beta_5 d_5 + \beta_6 d_6 + u_i$$

Where

Y_{1i} = Internally Generated Revenue for year 2012

Y_{2i} = wage bills for year 2012

d_{2i} = 1, if the state i is in the Southsouth

d_{2i} = 0, otherwise (any zone other than southsouth)

d_{3i} = 1, if the state i is in the southeast

d_{3i} = 0, otherwise

d_{4i}= 1, if the state i is in the North central

d_{4i}= 0, otherwise

d_{5i}= 1, if the state i is in the northwest

d_{5i}= 0, otherwise

d_{6i}=1, if the state i is in the northeast

d_{6i}= 0, otherwise

In this model, we have only qualitative regressors, taking the value of 1 if the observation belongs to a specific category and 0 if it belongs to any other category.

Now, taking the expectation of both sides, we obtain the following:

Average Y_i in southwest zone, the base group category which no dummy is assigned,

$$E(Y_i / d_{2i} = d_{3i} = d_{4i} = d_{5i} = d_6 = 0) = \beta_1 \text{.....} i$$

Average Y_i in the southsouth zone

$$E(Y_i / d_{2i}, d_{3i} = d_{4i} = d_{5i} = d_6 = 0) = \beta_1 + \beta_i \text{.....} ii$$

Average Y_i in the southeast zone

$$E(Y_i / d_{3i} = d_{2i} = d_{4i} = d_{5i} = d_6 = 0) = \beta_1 + \beta_i \text{.....} iii$$

Average Y_i in the north central

$$E(Y_i / d_{4i} = d_{2i} = d_{3i} = d_{5i} = d_6 = 0) = \beta_1 + \beta_i \text{.....} iv$$

Average Y_i in the northeast zone

$$E(Y_i / d_{5i} = d_{2i} = d_{3i} = d_{5i} = d_6 = 0) = \beta_1 + \beta_i \text{.....} v$$

Average Y_i in the northwest zone

$$E(Y_i / d_{6i} = d_{2i} = d_{3i} = d_{4i} = d_{5i} = 0) = \beta_1 + \beta_i \text{.....} vi$$

(The error term does not get included in the expectation values as it is assumed that it satisfies the usual OLS conditions, i.e., $E(U_i) = 0$)
To find out if the average internally Generated Revenue and wage bills of the Geopolitical zone are statistically different from each other (the comparison category), we have to find out if the slope coefficients of the regression result are statistically significant. For this, we need to consider the *p* values.

RESULT AND DISCUSSION

Using SPSS 15 to regress Internally Generated Revenue (IGR) and Wage bills on dummy variables (d_{2i}, d_{3i}, d_{4i}, d_{5i} and d_{6i}) the following results are obtained.

$$Y_{IGR} = 50.38 - 23.747d_2 - 42.905d_3 - 44.797d_4 - 42.723d_5 - 47.063d_6 + u_i$$

$$t = (3.019)(-1.051)(-1.714)(-1.983)(-1.955)(-2.083)$$

$$p - value = (.005)(.302)(.098)(.057)(.061)(.047)$$

Similarly,

$$Y_{Wagebills} = 44.06 + 8.007d_2 - 22.585d_3 - 16.443d_4 - 22.6d_5 - 23.443d_6 + u_i$$

$$t = (5.714)(0.767)(-1.953)(-1.575)(-2.195)(-2.245)$$

$$p - value = (.00)(.45)(.061)(.127)(.037)(.033)$$

The average Internally Generated Revenue and Wage bill in the southwest (base group) is the constant (50.38 and 44.06) of the model
From the expectation of the model, the following results (average internally generated revenue and wage bills of the six Geopolitical zones) were derived.

S/N	Geopolitical zones	Average IGR in billion	Average Wage bills in billion
1	Southwest	50.3800	44.0600
2	Southsouth	26.6333	52.0667
3	Southeast	7.4750	21.4750
4	North central	5.5833	27.6167
5	Northwest	7.6571	21.9000
6	northeast	3.3167	20.6167

From the results above, the estimated slope coefficient of IGR model for the southsout zone is not statistically significant as its *p* value is 30.2 percent; however, that of the southeast, north central, northeast and northwest zones are statistically significant at the 10% level as its

p values are 9.8, 5.7, 6.1 and 4.7 percent. Therefore, it can be concluded that the average Internally Generated Revenue in the southwest and southsouth zones are not statistically different from each other but the average Internally Generated Revenue in the Southwest, north central, northeast and northwest zones are statistically lower than that of the southwest zone.

Similarly, the estimated slope coefficient of Wage bill model for the southsout and north central zone are not statistically significant as its *p* value are 45 and 12.7 percent; however, that of the southeast, northeast and northwest zones are statistically significant at the 10% level as its *p* values are 6.1, 3.7, and 3.3 percent. Therefore, it can be concluded that the average wage bill in the southwest, southsouth and north central zones are not statistically different from each other but the average wage bill in the Southwest, northeast and northwest zones are statistically lower than that of the southwest zone. Comparing the revenue bases of the each Geopolitical zone to its wage bills (l.e salaries), apart from the southwest, the average wage bills of the southsouth zone is two times of its revenue, average wage bills of southeast and northwest zones are almost three times of its revenue, while that of north central and northeast are five times of its revenue.

CONCLUSION

From the analysis carried out, it was found that only southwest and southsouth are fairly strong to revenue bases, while that of southeast, north central, northeast and northwest are relatively very low even compare to the wage bills of the workers.

REFERENCES

Asha Sharma, Susan Garavaglia. "A SMART GUIDE TO DUMMY VARIABLES: FOUR APPLICATIONS AND A MACRO".

Gujarati, Damodar N (2003). *Basic econometrics*. McGraw Hill. p. 1002. ISBN 0-07-233542-4.

Draper, N.R.; Smith, H. (1998) *Applied Regression Analysis*, Wiley. ISBN 0-471-17082-8 (Chapter 14)

Wooldridge, Jeffrey M (2009). *Introductory econometrics: a modern approach*. Cengage Learning. p. 865. ISBN 0-324-58162-9.

Suits, Daniel B. (1957). "Use of Dummy Variables in Regression Equations". *Journal of the American Statistical Association* **52** (280): 548–551. JSTOR 2281705

John Fox, Applied Regression Analysis, Linear Models, and Related Methods, Sage Publications, 1997.