INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS

ISSN(print): 2643-9840, ISSN(online): 2643-9875

Volume 05 Issue 02 February 2022

DOI: 10.47191/ijmra/v5-i2-36, Impact Factor: 6.072

Page No. 482-487

The Effect of Categorizing a Continuous Independent Variable on ExplainedVariance

Denis Achung Uyanah[1], Henry Ojating[2]

[1,2]Department of Curriculum and Instructional Technology, Faculty of Education, Cross River University Of Technology Calabar, Nigeria

Abstract: Researchers, Particularly those in behavioral sciences, often categorize a continuous independent variable, in their research work. This is done for the main purpose of applying analysis of variance in testing their hypotheses. Such categorization is accompanied by loss of information which is reflected in a change in proportion of the variance in the dependent variable, accounted for, by the independent variable. To provide an empirical backing to this claim, a validated "teachers variables and the attainment of the goals of the Universal basic education scheme", was administered on a random sample of 640 secondary school teachers in Ogoja Education zone of Cross River State,Nigeria. The resulting data were analyzed using simple and multiple linear regressions. The independent variables were then categorized using sample mean and standard deviation. Analysis of variance was then applied. The proportion of the variance accounted for by the independent variables, individually and collectively were then compared using Fishers Z-transformation test. The results show that differences in explained variances do exist in favour of linear regression analysis, though not significant. The implication of the findings in research design is discussed.

KEY WORDS: Continuous Variable, Categorization, Explained Variance, Analysis of Variance, Factorial Design, Regression Analysis

INTRODUCTION

Analysis of variance (ANOVA) and regression analysis are two data analysis techniques that are very similar. This similarity arises from the fact that a regression model can be fitted into purely experimental data. Appealing as this may be, there are differences in the validating conditions, also called assumptions. These differences show-up in the results and decisions that are taken or that emanate therefrom.

Aczel and Sounderpandian (2006) and Anderson, Sweeney and Williams (1991) among others, stated that for analysis of variance to be validly Applied the dependent or response variable should be normally distributed in its population; the variance of the dependent variable should be the same for all sub-groups and that each of the sub-group must have been selected randomly and independently. In the ANOVA process, two independent estimates of the population variance, σ^2 are obtained. One estimate of σ^2 is obtained based on the differences between the treatment means ($\bar{x}i$) and the overall sample mean (μ). The other estimate of σ^2 is obtained based on the differences of observed values of the dependent variable within each treatment, from the corresponding treatment mean. The two estimates are then compared to determine whether the treatments means, assumed to have come from different population, are equal or not.

For regression analysis, the basic assumptions are: the error term in the regression model is a random variable whose mean is zero; the variance of the error term is the same for all values of the independent variable; the value of the error term for a given value of the independent variable is not related, in any form, to the value of the same error term for another value of the independent variable; the error term is a normally distributed random variable.

The similarities in assumptions are visible except one-the requirement that the mean of the error term is zero. All the others are equivalent. For example that the error term is a random variable whose distribution is normal is equivalent to the assumption that the dependent variable should be normally distributed in its population. This holds because the dependent variable is a linear function of the error term. So that if the error term is normally distributed, then the variable from which the error term was obtained, is equally normally distributed.

The Effect of Categorizing a Continuous Independent Variable on Explained Variance

Kerlinger and Pedhazur (1973) observed that researchers especially those in the behavioral sciences, are in the habit of categorizing a continuous independent variable for the purpose of using either independent sample t-test or analysis of variance. Although it may be valuable to conceptualize research design issues in this way, they maintained that it is quite inappropriate to analyze them that way. So such procedures and approaches throw away useful information.

Kemeny, Snell and Thompson (1966) observed that when one dichotomizes a continuous variable that can take a range of numerical values, one loses considerable variance. This can be extended to a situation where the continuous independent variable is partitioned into three or more categories, whether using the observed mean and standard deviation, theoretical mean and standard deviation inter-person correlation or factor analysis (Uyanah, 2014). This may mean that inter-variable correlations are lowered to a level of non-significance, when in actual fact; the tested relationships may be significant (Kerlinger & Pedhazur, 1973). Those who adopt this approach, very often force the analysis of their research data on the procrustean bed of the elegance of the preferred method.

It should be noted that when a continuous independent variable is partitioned, the partitions are subsets of sets that are disjoint or mutually exclusive and exhaustive. The separate partitions have no quantitative meaning except that of qualitative difference. Moreover, the numbers assigned to these partitions are really labels that do not have numerical meaning. The level or scale of measurement has been reduced to nominal level, so that they cannot be ordered, added or subtracted, strictly speaking (Kerlinger & Pedhazur, 1973).

Regression analysis theoretically is and remains superior compared to analysis of variance in situations where:

1. The independent variable is continuously measured
2. There are two or more independent variables that are mixtures of both continuous and categorical
3. When the number of observations per cell are both unequal and disproportionate
4. When studying trends in a dependent variable, whether linear, quadratic, cubic etc.

According to Kerlinger & Pedhazur (1973), this list is only illustrative and not exhaustive. They argued that analysis of variance in such situations should be a stepping stone towards a conclusion that there exists a linear relationship. This, they continued, can be done by carrying out both regression analysis and analysis of variance, and testing the departure from linearity for significant. If there is departure from linearity, the between treatments sum of squares will always be larger than the sum of squares due to regression. This is where the superiority of regression analysis is, because further analysis can be done by fitting a polynomial regression model that may explain a higher amount of the variation in the dependent variable. If the regression sum of squares is higher than the between treatments means sum of squares, then the regression analysis is most appropriate.

This extent is traditionally not reached by researchers. Some do not even believe or are aware that regression in some or all situations is superior to ANOVA. Even those who believe, do so intuitively. They are some situations where high ranking academics, supervising post graduate students insist that, a continuous independent variable should be categorized and ANOVA applied. This makes the provision of empirical evidence imperative. This is what this study sought to do-to compare the proportion of the variance in the dependent variable accounted for by the independent variable when it is categorized and ANOVA applied, with the sum of squares that are obtained from regression analysis.

METHODOLOGY

Ex-post-factor research design was adopted for the study, because the independent variables were not manipulated. The variables had already interacted and produced their effect now observed as UBE goal attainment. A 47-item instrument tagged “Teachers Variables and UBE goal attainment questionnaire” was used for the study.

The instrument was previously developed and validated by Unwanede (2016). It was designed to measure four teachers’ variables-perceived remuneration, capacity building, teacher-principal relationship and teacher-learner relationship (eight items each) with UBE goal attainment (15 items). All the items were built on a four (4) point modified Likert scale. The Cronbach alpha reliability estimates for the five substances are presented in Table 1.

The Effect of Categorizing a Continuous Independent Variable on Explained Variance

Table 1. Cronbach Alpha reliability estimate for the five (5) sub-scales (variables)

Name of sub-scale	**No. of**		**Std.**	**Std.**	**Sum of**	**Cronbach**
	items	**mean**	**dev.**	**error**	**item var.**	**Alpha**
Trs. Remuneration	8	17.353	5.700	.225	9.680	.802*
Trs. Capacity building	8	20.698	4.042	.715	8.446	.681*
Prin.-Trs. Relationship	8	23.025	3.726	.186	7.119	.778*
Trs.-learners relationship	8	22.252	4.145	.147	7.802	.690*
UBE goal attainment	15	38.168	6.906	.297	12.450	.810*

The results in Table 1 show that all the reliability coefficients are significantly higher than the .500 recommended by Nunnally (1978). The instrument was therefore considered useable.

The instrument was administered on a random sample of 640 secondary school teachers in Ogoja Educational Zone of Cross River State, Nigeria. Their responses were weighted as follows: 4 for Strongly Agree (SA), 3 for Agree (A), 2 for Disagree (D) and one (1) for Strongly Disagree (SD); for positively worded statements and reversed for negatively worded items. Theses weights were added for each variable per respondent. The independent variables were categorized into three partitions identified as high (x > $\bar{x}$ + s) moderate ($\bar{x}$ - s ≤ × ≤ $\bar{x}$ + s) and low (x < $\bar{x}$ – s). Simple and multiple linear regression analyzes were used to analyze for the individual and collective influence of the independent variable(s) on UBE goal attainment.

The analyses were repeated using one-way and four-ways ANOVA. The proportion of the total variance in the dependent variable accounted for by the independent variables were obtained and correspondingly compared, using Fishers' Z-transformation method. Decisions about the significance of the results were taken by comparing the P-values associated with the computed test statistics to the chosen level of significance (.05). So that results were said to be significant if the observed P-value was less than .05 and not significant if the P-value was greater than .05.

RESULTS

The descriptive statistics, mean standard deviation, standard error, minimum and maximum- of the five research variables are presented in Table 2.

Table 2. Descriptive statistics of the five research variables

Name of research variable		**Std.**	**Std.**	**Mini**	**Maxi**
	mean	**dev.**	**error**	**mum**	**mum**
Trs. Remuneration	17.335	5.695	.225	8	32
Trs. Capacity building	20.998	4.429	.175	9	32
Prin.-trs. Relationship	23.502	4.722	.187	8	32
Trs.-learners relationship	23.225	4.415	.174	9	32
UBE goal attainment	39.816	7.069	.279	16	60

N=640

The results in Table 2 show that only the mean perceived teacher's remuneration ($\bar{x}$ = 17.353) is less than the expected mean of (μ = 20.00). All the other mean values are greater. These differences were not tested because it fell outside the scope of this study.

The results of simple and multiple linear regression analysis as well as the corresponding one-way and four-ways ANOVA, are presented in Table 3

The Effect of Categorizing a Continuous Independent Variable on Explained Variance

Table 3. Linear regression and ANOVA for the influence of teacher's variables on UBE goal attainment

Teachers' variable	Nature of variance	Linear regression	Analysis of variance
Teacher Remuneration	Explained	235.290	182.917
	Error	31746.988	31799.361
	Total	31982.278	31982.361
		a=39.101 b=.041	R= .077
		R=.086, R^2=.007	R^2= .006
	Remark	Not significant	Not significant
Teachers' Capacity building	Explained	2890.271	1610.762
	Error	29173.007	30371.516
	Total	31982.278	31982.278
		a=35.587, b=.201	R = .224
		R=.296, R^2=.088	R^2= .053
	Remark	Significant	Significant
Principal-teacher's Relationship	Explained	1494.611	383.416
	Error	30487.667	31598.862
	Total	31982.278	31982.278
		a=32.210, b=.324	R= .110
		R=.216, R^2=.047	R^2= .012
	Remark	Significant	Significant
Teacher's-learners relationship	Explained	2717.920	1559.443
	Error	29264.358	30422.855
	Total	31982.278	31982.278
		a=28.977, b=.467	R= .221
		R=.292, R^2=.085	R^2= .049
	Remark	Significant	Significant
All four variable together	Explained	3461.864	10557.227
	Error	28520.414	21425.051
	Total	31982.278	31982.278
		R= .329	R= .574
		R^2 = .108	R^2= .330
	Remark	Significant	Significant

The results in Table 3 reveal that for the influence of teachers' remuneration, capacity building, teacher-principals' relationship and teacher-learners' relationship, taken individually, the decisions taken using regression analysis results were the same with the decisions taken using ANOVA results. Even when all the independent variables were taken, the decision taken using the multiple regression results was the same with those using univariate ANOVA.

In-terms of the proportion of the total variance accounted for (R^2) the values from regression analysis were higher than those from ANOVA when the independent variables were considered individually. The results specifically show that for perceived teachers' remuneration, regression approach accounted for .7% of the total variance while ANOVA approach accounted for .6%. For perceived teachers' capacity building, regression approach explained 8.8% of the total variance while ANOVA explained 5.3%. In the case of principal-teachers' relationship, regression analysis explained 4.7% while ANOVA approach explained only 1.2%. When the influence of teachers-learners relationship was analyzed for, the regression approach explained 8.5% of the total variance while ANOVA approach accounted for 4.9% only. When they were taken collectively, the ANOVA R^2 (33.0%) was higher than that from multiple linear regression analysis (10.8%).

To test for the significance of these differences in R-squared, the values were converted to R by taking square root. The Fishers' Z-transformation test was then applied. The results are presented in Table 4.

Table 4. Fisher's Z-transformation comparison of proportion of variance explained by ANOVA and linear regression analysis

Independent variable	Linear regression (R_1)	ANOVA (R_2)	Z of (R_1)	Z of (R_2)	Z
Trs. Remuneration	.086	.077	.090	.075	.268
Trs. Capacity building	.296	.224	.310	.229	1.446
Prin.-trs. Relationship	.216	.110	.218	.110	1.957
Trs.-learners relationship	.292	.221	.304	.224	1.450
UBE goal attainment	.329	.574	.343	.655	56.549*

significant at .05 level. Critical Z= ± 1.96

From Table 4, only the difference in R-squared between multiple regression and univariate ANOVA was significant (Z=56.549, Critical Z=1.96, α = .05). All the other comparisons were not significant because all the observed z-values (.268, 1.446, 1.957, & 1.450) are less than the critical z-value (1.96).

DISCUSSION

The results that regression analysis is superior to ANOVA when the independent variable is continuous, is very interesting. Though the differences in the proportion of variance accounted for by the independent variable were not significant, they are however a pointer to the fact that there exist probabilities that at some point the differences may become significant. These results agree with the position taken by Kerlinger and Pedhazur (1973) that when a continuous independent variable is partitioned for the purpose of using either independent t-test or analysis of variance, some amount of variance is lost.

As it stands and with respect to the simple linear regression compared with one-way ANOVA, the conclusion is that the linear regression model fits the data more than the ANOVA model. This follows from position of Kerlinger and Pedhazur (1973) that when the between groups or treatment sum of squares is larger than the regression sum of squares, there may be significant departure from linearity. Now that the reverse is the case, it follows that linearity is the most appropriate.

The results that show that when all the four variables were partitioned and four-way ANOVA applied, the between groups sum of squares is larger than the multiple regression sum of square is interesting also. It should be noted that this does not mean that the ANOVA approach is superior. It shows that there seem to be a significant departure from linearity. If the study must be taken to a logical conclusion, then non-linear components of the regression must be added. The end-point may be the point where $SS_{reg} \geq SS_{between\ groups}$.

These results clearly show that it is inappropriate to partition a continuous independent variable, apply ANOVA and stop there. The best decision is to go straight to regression analysis or use the ANOVA approach only as a step towards obtaining the most adequate regression model.

REFERENCES

1) Aczel, A.D. & Sounderpandian, J. (2006). *Complete statistics* (6th ed.). New York: McGraw Hill.
2) Anderson, D.R. Sweeney, D.J. & Williams, J.A. (1991). *Introduction to statistics*: *concepts and applications* (2nd ed.) New York: West publishing company.
3) Kerlinger, F.N. & Pedhazur, E.J. (1973). *Multiple regression in behavioral research.* New York: Holt, Rine heart and Winston.
4) Kemeny, J.G., Snell, J.L & Thomason, G.L. (1966). *Introduction to finite mathematics.* (2nd ed.). Engle wood cliffs, N.J.: Prentice-hall.
5) Nunnally, J.C. (1978). *Psychometric theory* (3rd ed.). New York: Mc-Graw-Hill.
6) Unwanede, U.A. (2016). *The influence of teacher's Variables on Universal Basic Education (UBE) goal attainment in Ogoja Education Zone of Cross River State, Nigeria.* A master's thesis presented to the Post Graduate School, cross River University of Technology, (CRUTECH) Calabar, Nigeria.

7) Uyanah, D.A. (2014). *Desirable variance maximization in research resigns*. A paper presented at the 16th annual conference of association of Educational Researchers and Evaluators Nigeria (ASEREN). 14th-20th July. University of Calabar, Calabar, Nigeria.