GSJ: Volume 11, Issue 1, January 2023
ISSN 2320-9186

GSJ: Volume 11, Issue 1, January 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

STATISTICAL REPERCUSSION OF REGRESSION ASSUMPTION VIOLATION ON MODEL APTNESS

OMOKARO, B. E. & AKPOJARO, O. O
DEPARTMENT OF STATISTICS, DELTA STATE POLYTECHNIC, OTEFE-OGHARA
owenssmith@gmail.com, +2347067000466

ABSTRACT

The goal of this study is to assess the statistical effects of breaking the regression assumptions. Regression modeling was used in the study with an emphasis on testing the residual and normalcy assumptions. In this study's data analysis, we used two sets of data: one original and the other a contaminated copy of the original. Regression analysis was used to assess the study's findings in order to validate the regression assumptions, notably the residual analysis and the test for normality. The study result found that both models on the different set of data indicated that they are adequate. Further analysis revealed that the model with original values satisfied the regression assumption of normality and residual and the coefficient estimate were higher compared to the second model where the assumptions were violated. Although the model appears to be appropriate, the parameter estimates were small, and any inference made from the tainted dataset may be false. Therefore, it is crucial to validate the regression model's assumptions through empirical research because doing so could have disastrous consequences.

Keywords: Model aptness, regression modeling, model assumption, statistical analysis, parsimonious model

Introduction

The proper analysis of data is crucial for making judgments on issues affecting economic, business, social, academic, and general organizations. Many statistical methods that are common to the subject of econometrics, such the regression methodology, must be properly researched in order to determine how most situations are thought to be affected by particular characteristics. Numerous areas of human endeavor have greatly benefited from the regression theory during the years since its development. As a result, this study will give readers a fundamental understanding of the notion of regression and how it may be used to model economic variables. Ogbogbo et al (2014) carefully explained that regression is the study of relationship among variables. The procedure of regression analysis involves formulating a mathematical equation to estimate the value of one variable (the dependent variable, Y) given the value(s) of one or more quantitative variable(s), (the independent variable(s), X). They further buttressed that there are various methods for estimating the coefficients of a regression model, this includes; freehand method, the least square method and the matrix method.

The fitted models must unquestionably meet the regression assumptions of independence of the data, nonlinearity of the explanatory variables, and normal distribution of the errors in order for the analyses to be considered acceptable. Invalid statistical inferences may result when certain assumptions of regression analysis are not met, since they may result in biased coefficient estimates or very large standard errors for the regression coefficients. They proposed that evaluating a regression model's fit, or how well it describes the observed data, is a crucial step in determining if it is acceptable. The conclusions generated from the model without such an investigation may be deceptive or even wholly erroneous.

Frost (2020) summarized that in a nutshell, linear model should produce residuals that have a mean of zero, have a constant variance, and are not correlated with themselves or other variables. If these assumptions hold true, the OLS procedure creates the best possible estimates. In statistics, estimators that produce unbiased estimates that have the smallest variance are referred to as being “efficient.” Efficiency is a statistical concept that compares the quality of the estimates calculated by different procedures while holding the sample size constant. OLS is the most efficient linear regression estimator when the assumptions hold true.

The application of regression analysis have gained wide acceptance across many field of research around the globe and this incorporates the application by both statisticians and non-statisticians who may or may not have possessed indept foundation knowledge of the technique. The lack of conceptual understanding brings about violation of assumptions guiding the appropriate utility of the model in researches. Parameter estimation in a regression model requires the assumption that the error terms are uncorrelated with mean zero and constant variance. Also, test of hypothesis and confidence interval construction demands that errors be normally distributed. However, where these assumptions are violated; the entire analysis and subsequent inference will be misleading. In light of this, this study seeks to evaluate the relevance of regression model assumption validation to researchers and students.

Review of related literature

Wolfram (2011) stated that the misconception that the normality assumption applies to the response and/or predictor variables is problematic in that there are certainly situations where the response and/or predictors are not normally distributed, but a normal distribution for the errors is still plausible. As one example, dichotomous predictors are often used in multiple regression; although such predictors are clearly not normally distributed, the errors of regression models using dichotomous predictors may still be normally distributed, allowing for trustworthy inferences. Furthermore, dichotomous variables that are particularly strong predictors of a response variable may induce a bimodality to the marginal distribution of the response variable, even if the errors are normally distributed. This is one situation in which neither predictor nor response variable has a normal distribution, despite the model errors being normally distributed.

Hoekstra et al. (2012) was of the opinion that some researchers did not check assumptions for certain reasons. They list unfamiliarity with either the fact that the model rests on the assumption or with how to check the assumption as the top two reasons. As explained, incorrect dealing with the assumptions could lead to severe problems regarding the validity and power of the results. We focused on four assumptions that were not highly robust to violations, or easily dealt with through design of the study, that researchers could easily check and deal with, and that, in our opinion, appear to carry substantial benefits.

Tariq et al. (2016) conducted an empirical test of the consequences and solution of OLS assumption violation specifically focusing on autocorrelation and heteroscedasticity problems in data. For this purpose, the relationship of Gross Domestic Product with the inflation, exchange rate and interest rate has been examined for Pakistan by using data from 1973-2008. Hodric-Prescott filter method has been used for making the data stationary. For obtaining reliable results and maintaining the properties of the coefficients appropriate consideration of the serial correlation and heteroscedasticity assumptions cannot be underestimated.

Opara and Jude (2021) examined the effect of non-normal error distribution on simple linear regression versus its nonparametric equivalent. The error term for normality proved that it is not from a normal population using Ryan-Joiner, which violates the major assumption of simple linear regression. Hence, estimating its slope becomes immaterial and any inference drawn from the OLS won’t be reliable. Since, there is no need of employing the technique, due to its poor performance in the presence of error non-normality, then a feasible alternative technique which performs consistently and robust to non-normality residual is required. The simulation study conducted in this study suggested that the nonparametric Theil’s simple linear regression is an alternative to OLS when there is existence of non-normal error in a data set. The study recommended among others that further studies on simple linear regression should ensure that the underlying assumptions of OLS are fulfilled before estimation; otherwise its non-parametric equivalent should be employed, but if the researcher must continue with OLS after failure of assumption, then outliers should be checked and if detected, should be removed and re-examine the underlying assumptions.

Methodology

The utilization of published reports as a secondary source of data gathering was used in this study. In this study, data from incidents of violence against women recorded over a five-year period (2017–2021) by the UN Entity for Gender Equality and Empowerment of Women was analyzed. The two-way analysis of variance (ANOVA) and Fishers Least Significant Difference

(Fishers LSD) will be used to analyze the data. The Fishers LSD will be used to determine which continent has a higher incidence of domestic violence against women, and the two-way ANOVA will be used to see whether there is a significant difference in the cases of domestic violence across the countries.

Model Specification

The study employed the method of multiple regression analysis in examining the variables of interest. In this study, the multiple regression analysis was employed. A multiple regression analysis is a statistical study of how a dependent variable or a variable that can be predicted is influenced by the changes in two or more explanatory variables.

Generally, the multiple regression model is of the form;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_{ij} \qquad (1)$$

The above model will be estimated using the matrix approach. Suppose that there are *p* independent variables and *n* observations ($x_{11}, x_{12}, \ldots, x_{1p}, y_i$), $i = 1, 2, \ldots, n$ and the model relating the independent variables to the dependent variable. Given the value of y denoted by y_i the system of the above equation becomes:

$$\begin{array}{lcl} y_1 & = & \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \varepsilon_1 \\ y_1 & = & \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \varepsilon_2 \\ \vdots & \vdots & \vdots \quad \vdots \quad \vdots \quad \vdots \\ y_1 & = & \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \varepsilon_n \end{array} \qquad (2)$$

Expressing the above system in matrix form we obtain as follows;

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} \cdots & X_{1p} \\ 1 & X_{21} & X_{22} \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The model is a system of n equations that can be expressed in matrix notation as defined:

$$y = X\beta + \varepsilon \qquad (3)$$

We desire to obtain the vector of least squares estimator, $\hat{\beta}$, that minimizes the error sums of squares (SSE).

Using equation 3, make ε the subject of formular

$$\varepsilon = y - X\beta \qquad (4)$$

The sum of squares of SSE

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) \qquad (5)$$

The least squares estimator $\hat{\beta}$ is the solution for β in the equations

$$\frac{\partial L}{\partial \beta} = 0$$

$$\varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

$$y'y - y'X\beta - X'\beta' y + X'\beta' X\beta \qquad (6)$$

$$y'y - y'X\beta - y'X\beta + X'\beta' X\beta \qquad (7)$$

$$\frac{\partial \varepsilon' \varepsilon}{\partial \beta} = 0 \to -y'X - y'X + 2X'X\beta$$

$$-2X'y + 2X'X\beta = 0$$

$$2X'X\beta = 2X'y$$

$$X'X\beta = X'y$$

$$\beta = \frac{X'y}{X'X}$$

GSJ: Volume 11, Issue 1, January 2023
ISSN 2320-9186

$$\hat{\beta} = (X'X)^{-'}X'Y \qquad (8)$$

$$\hat{\beta} = (X'X)^{-'}X'Y = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

Necessary assumptions and problems in the regression model

One of the most powerful and frequently referred and used statistical tool in the entire spectrum of data analysis. It is often mis-used by many practitioners. The concept of regression analysis is built on certain assumptions or decides to deliberately ignore them. It is important to note that when these assumptions are ignored. The results of the analysis might be false, misleading, resulting in wrong statistical inference.

These are basic assumptions about the error term.

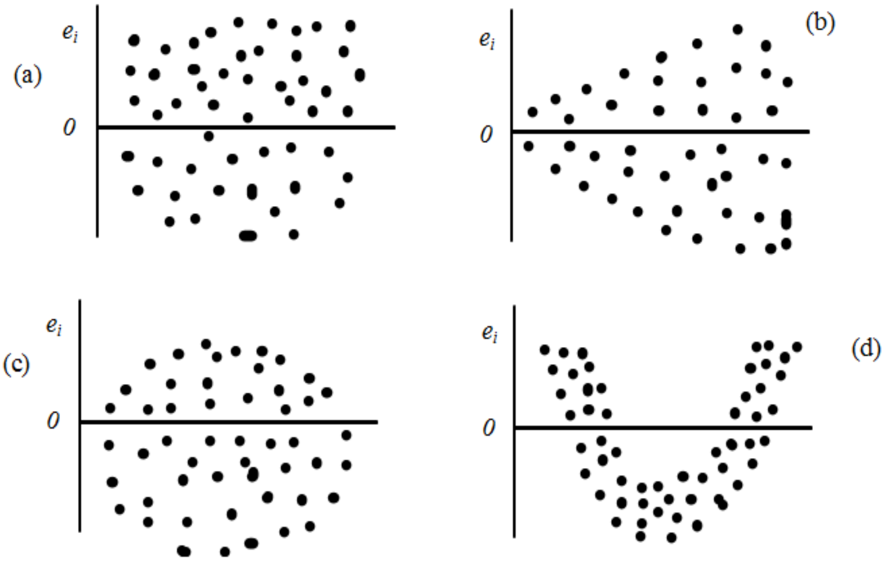
- The error term (e) is a random variable with mean equal zero (i.e. E(e) = 0)
- The variance of the e, denoted by σ^2, is the same for all values of X (constant variance assumption)
- The values of e are independent
- The error term, e, is a normally distributed random variable.

1. The mean is zero and the variance of the residuals is constant

Multiple linear regression assumes that the amount of error in the residuals is similar at each point of the linear model. Parameter estimation in a regression model requires the assumption that the error terms are uncorrelated with mean zero and constant variance. Thus, residual from a regression model is defined:

$e_i = y_i - \hat{y}_i$, i = 1, 2, ..., n where y_i is the observe d value (actual observation) and $\hat{y}_i$ is the corresponding fitted estimated value from the regression model.

The resulting graphical plot is given below:



Residual analysis helps us understand whether the model assumptions are violated or not. One useful way to analyze residuals is to express them versus various criteria. The resulting plots are called residual plot. To construct a residual plot, we compute the residual for each of the observed *y* values. To validate the regression assumptions, we make residual plot against:

a) Values of the independent variables
b) Values of $\hat{y}_i$ the predicted value of the dependent variable
c) The time order in which the data have been observed.

The residual plot (*a*) shows a random pattern of clustering of the plotted points representing an ideal situation (homoscedastic) while patterns *b*, *c*, and *d* represent abnormalities suggesting a violation of the classical assumptions of regression analysis.

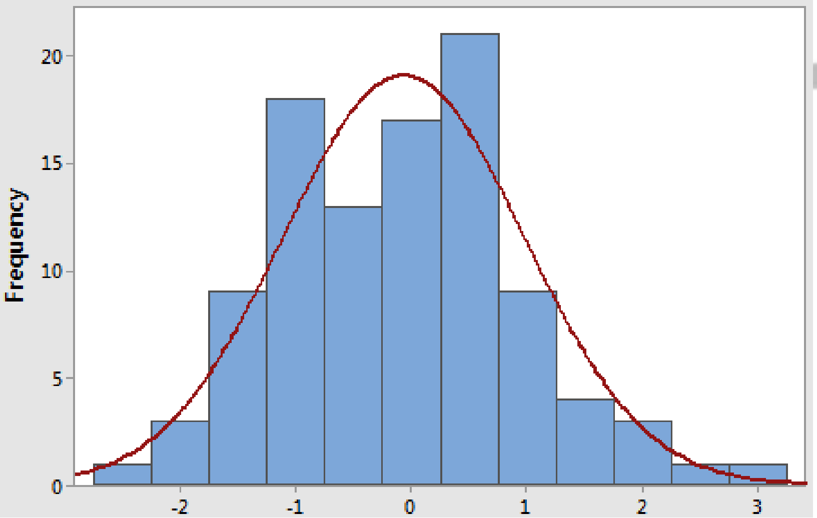
2. Independence of observation

The model assumes that the observations should be independent of one another. Simply put, the model assumes that the values of residuals are independent. To test for this assumption, we use the Durbin Watson statistic. The test will show values from 0 to 4, where a value of 0 to 2 shows positive autocorrelation, and values from 2 to 4 show negative autocorrelation. The mid-point, i.e., a value of 2, shows that there is no autocorrelation.

3. Multivariate normality

Multivariate normality occurs when residuals are normally distributed. To test this assumption, look at how the values of residuals are distributed. It can also be tested using two main methods, i.e., a histogram with a superimposed normal curve or the Normal Probability Plot method.

Histograms are particularly problematic when you have a small sample size because its appearance depends on the number of data points and the number of bars.



Results

The data for the study were analysed using SPSS statistical software and the result are presented in tables and figures as shown below. In order for us to provide a more clearer explanation about the occurrence and happening in the research problem, the study therefore utilized both descriptive and inferential method in assessing the presented data through the aid of the SPSS.

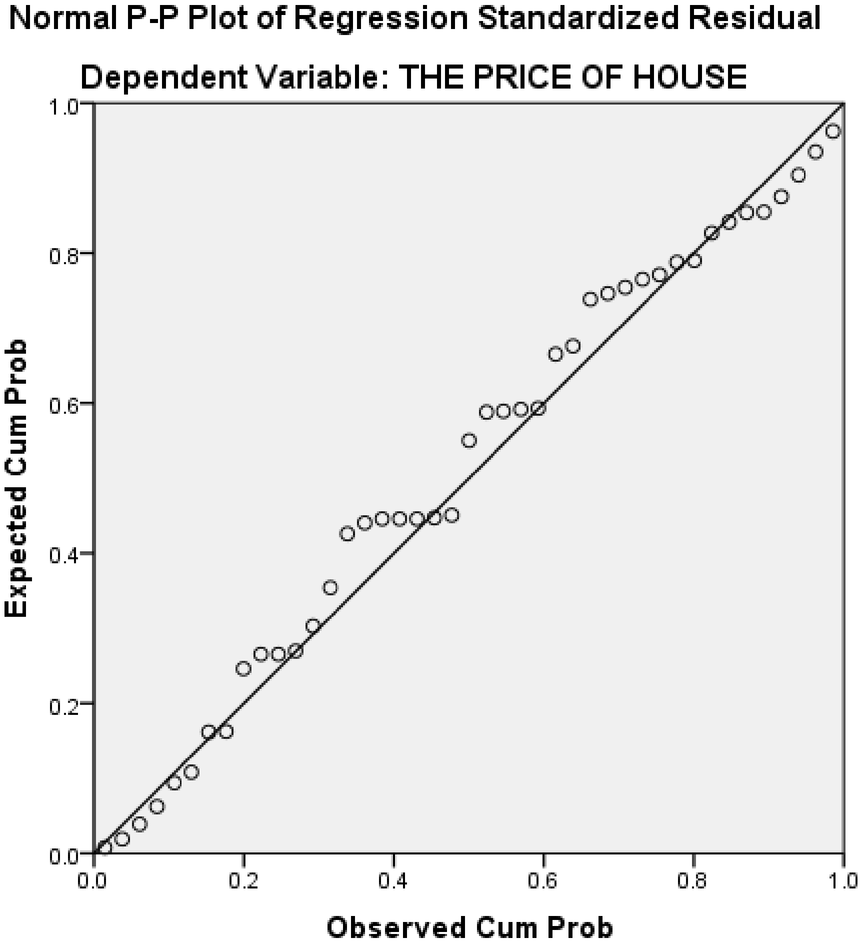
Table 1: Descriptive statistics

	N	Minimum	Maximum	Mean	Std. Deviation
THE PRICE OF HOUSE	43	45000.00	120000.00	93116.2791	17963.20220
NUMBER OF TENANTS	43	3.00	24.00	9.0233	3.97305
AGE OF THE HOUSE	43	1.00	8.00	3.2558	2.05974
Valid N (listwise)	43				

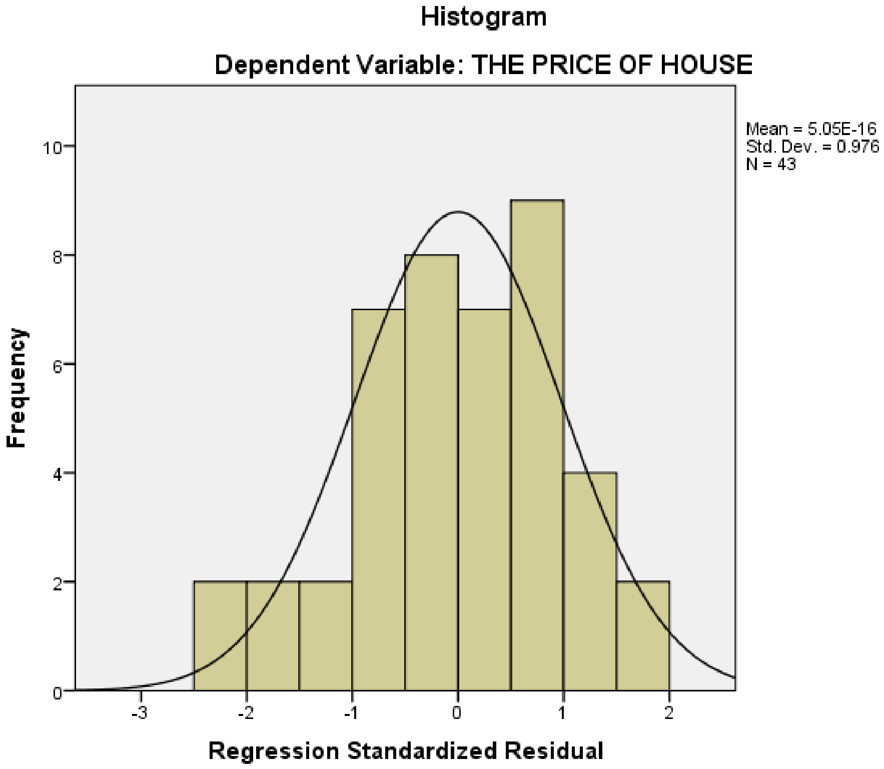
The result in table 4.2 provides the information that the average/mean value of advertisement expenditure and sales volume after advert. It was also shown that on the average, the price of rentage in Otefe housing is about N93,000, an average of 9 tenants resides in a compound in the houses on rent in Otefe and the study also revealed that on the average houses in Otefe that were studied are of age of 3 years old.

Verification of normality assumption

Using Normal P-P Plot



Ideally, for a normally distributed random error term plot will produce a plot where the datapoints cluster closely around the normal line as shown in the above. If your data is not normal, the little circles will not follow the normality line. Sometimes, there is a little bit of deviation, such as the figure all above but we can assume normality as long as there are no drastic deviations.



The histogram plot also suggest a normality of the data since the curve is asymptotic with a single in the middle possessing a bell shape like structure.

Verification of residual assumption

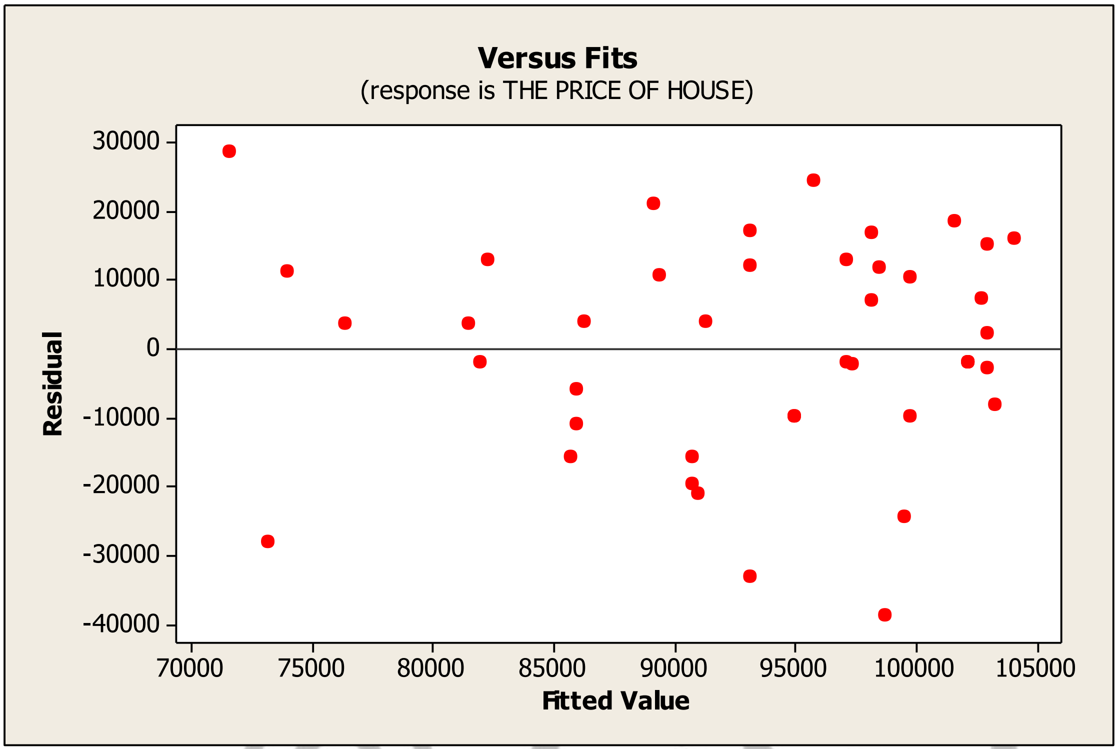
Table 2: residual analysis

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	71607.6016	104054.0469	93116.2791	8954.76176	43
Residual	-38731.08594	28392.39648	.00000	15572.05431	43
Std. Predicted Value	-2.402	1.221	.000	1.000	43
Std. Residual	-2.427	1.779	.000	.976	43

a. Dependent Variable: THE PRICE OF HOUSE

From the study, result indicated that the mean of the residual is zero (0) and a standard deviation of 0.976 which is approximately 1.0 which is a primal indication that the data fits the assumption of the use of regression model since the original assumption states that the mean of the residual is zero with a standard deviation of 1, but a clear understanding and verification of this situation will be provided by the result of the graph below:



The study also indicated in figure describing the residual and predicted value plot, the data is random and thus with the resulting values of the mean of the residual, it is therefore convenient to say that the result of this study is very significant and thus the regression model can be fitted.

Table 4: Regression parameter estimates

Table 3: Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	109377.012	6896.956		15.859	.000
	NUMBER OF TENANTS	-267.461	624.563	-.059	-.428	.671
	AGE OF THE HOUSE	-4253.120	1204.726	-.488	-3.530	.001

a. Dependent Variable: THE PRICE OF HOUSE

In the above result, the analysis revealed that averagely, the price of rentage of resident house in the community will be about N109,000 if the age of the house and the number of tenants are not considered. However, the price of the rent will significantly reduce with about N4,200 for every increase in the age of the house. This implies that the older the building the lesser the range cost associated with it since people usually prefer to rent new and flashy buildings. The study also revealed that the impact of the number of tenants in the building have not significant effect on the price of rentage.

MODEL HYPOTHESIS

H_0: The estimated regression model is inadequate

H_I: The estimated regression model is adequate

Level of significance = 0.05

Table 4: ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3367885839.879	2	1683942919.939	6.614	.003[b]
	Residual	10184532764.772	40	254613319.119		
	Total	13552418604.651	42			

a. Dependent Variable: THE PRICE OF HOUSE

b. Predictors: (Constant), AGE OF THE HOUSE, NUMBER OF TENANTS

DECISION RULE

Reject Ho if the p-value < 0.05 level of significance

DECISION

Since the p-value (0.003) is less than the level of significance, we accept H_1 and conclude that the model is adequate. This implies that the variables included in the model are capable of influencing the price of rentage.

Simulation study

The data for the study were altered for further diagnostic effect, the data are presented as follows:

Table 5: Simulated data Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	96756.968	6483.206		14.924	.000		
	NUMBER OF TENANTS	1087.195	910.458	.240	1.194	.239	.521	1.920
	AGE OF HOUSE	-1921.533	743.559	-.520	-2.584	.014	.521	1.920

a. Dependent Variable: PRICE OF RENTAGE

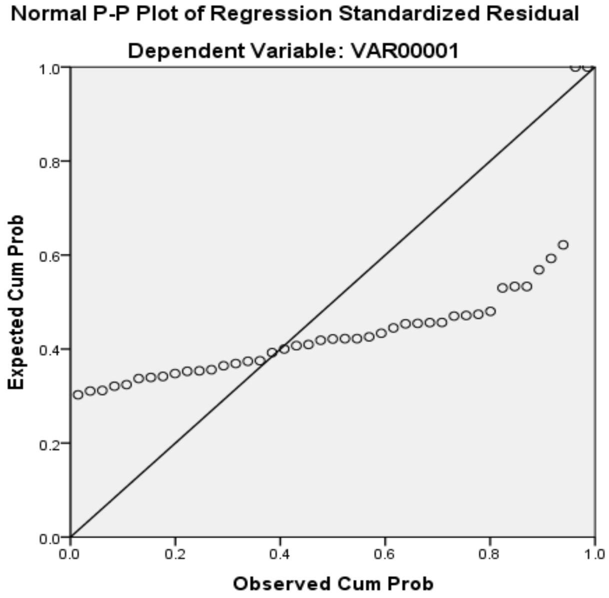
The model also signified that the age of house is significant in predicting the price of rentage in the community, though the variable parameter estimates changes. The result presented in table

4.4 with the original data indicated that the age of house will reduce the price of rentage with about N4,000 but in this case, it changes with only about N1,900.

Verification of normality assumption

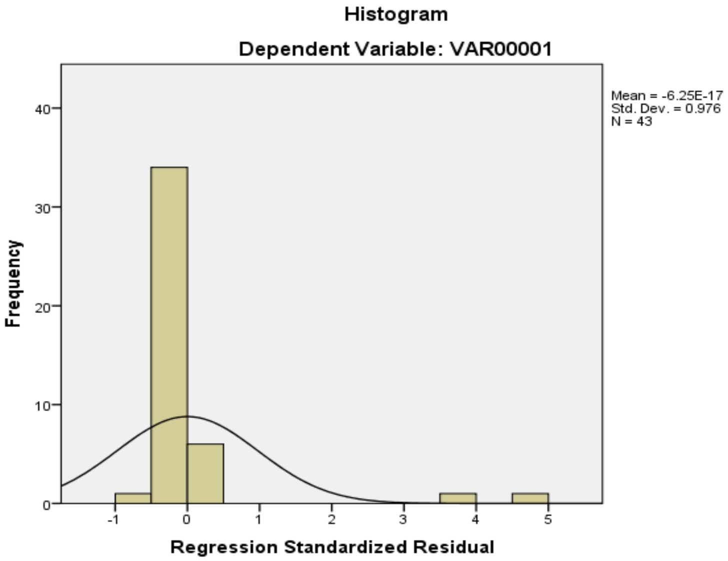
Upon examining the assumption of the model, the following were reported

Using Normal P-P Plot



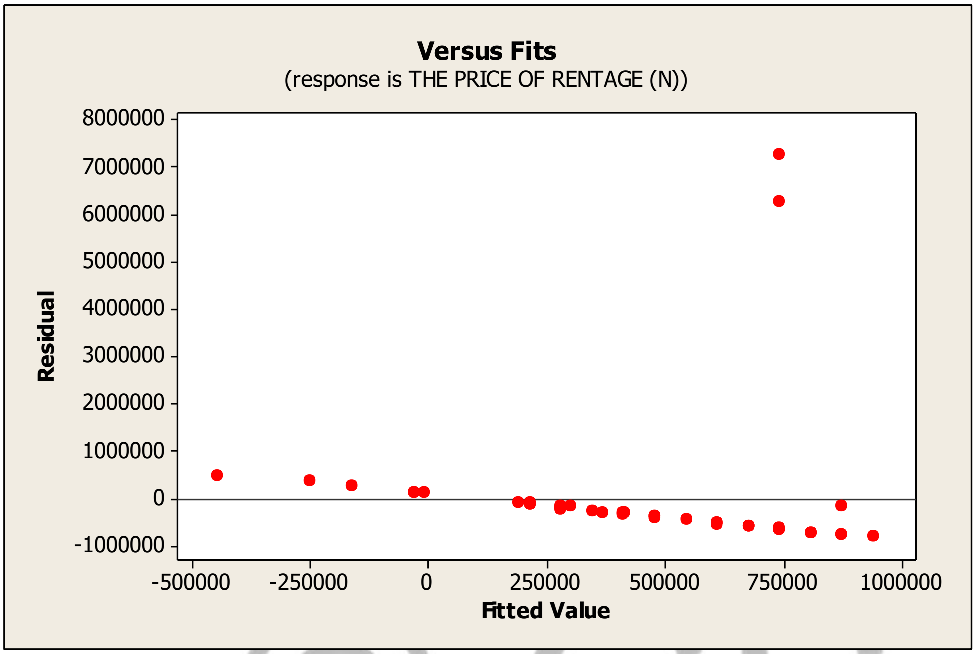
Ideally, for a normally distributed random error term plot will produce a plot where the datapoints cluster closely around the normal line but in this case there is a huge variation as the data deviated to forming a horizontal line instead of diagonal indicating that the normality assumption is violated.

Using histogram



The histogram plot could not produce a bell shape, thus indicating that the normality assumption is violated.

Residual analysis



The above graph it was shown that that the fitted values did cluster around zero, showing that the mean of the error term is zero, but the constant variance assumption was violated since most the data cluster at the edge of the plot and the presence of the two outlying data points.
In essence, despite that the model proved to be adequate when the original values are altered, the new model could not sustain the satisfaction of the regression model assumptions. This resulted to the effect in the sake of prediction since the model parameters could not also remain the same and cannot be relied upon.

Conclusion

In this study's data analysis, we used two sets of data: one original and the other a contaminated copy of the original. Regression analysis was used to assess the study's findings in order to validate the regression assumptions, notably the residual analysis and the test for normality. According to the study's findings, both models on the various sets of data revealed that they are sufficient. The model with original values satisfied the regression assumption of normality, and the coefficient estimate was larger compared to the second model, where the assumptions were broken, according to further research. Although the model appears to be appropriate, the parameter estimates were small, and any inference made from the tainted dataset may be false. Therefore, it is crucial to validate the regression model's assumptions through empirical research because doing so could have disastrous consequences. The study's findings led to the advice that, in order to make efficient decisions for statistical inference, researchers should consider validating the underlying assumptions when building effective models in empirical research.

GSJ: Volume 11, Issue 1, January 2023
ISSN 2320-9186

References

Akpojaro, O.O & Onwubuya, M.N, (2020): A model of some Nigeria socio-economic problems: multiple regression approach, global scientific journal, vol 8, issue 1 pg 2867

Frost, E. E. (2020): Multiple Linear Regression Models for Estimating Microbial Load in a Drinking Water Source Case from the Glomma River, Norway:, available at www.core.ac.uk/download/pdf/3088960

Hoekstra, R., et al. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? Frontiers in Psychology, 3. doi:10.3389/fpsyg.2012.00137

Ogbogbo, G. O., et al. (2014). Essentials of Descriptive Statistics. Robert Press and Publishing Company

Opara. J. & Jude, E. (2021). Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research & Evaluation 8(2):1–9.

Tariq, B., Sarkar, H. & Midi, K. (2016). Beyond Linearity by Default: Generalized Additive Models. American Journal of Political Science. 42 (2): p 596-627.

Wolfram, W. R. (2011): Advertising Expenditures as an Economic Stabilizer. *Quarterly Review of Economics & Business*, pp. 7-18