Journal of the Nigerian Association of Mathematical Physics
Volume **22 (November, 2012), pp 249 – 256**
© J. of NAMP

Assessing the Detection and Correction of Model Violations Using Residuals in a Linear Regression Diagnostics

Osemeke Reuben. F. and Ehiwario, J.C.
Department Of Mathematics
College Of Education, Agbor

Abstract

In ordinary linear regression, graphical diagnostics and numerical test were used to detect and correct the model violations of regression assumption. Residual plots against the predictor or predicted ($\hat{y}_i$) were used to show error violations of heteroscedasticity, autocorrelation, outliers, clustering of data points and nonlinearity. Non-normality was diagnosed using letter value display. The error violations were corrected by plausible trial and error transformation of the variables. Analysis of residuals after the correction were improved upon as shown in the coefficient of determination (r^2),multiple r, p value < 0.05, Durbin Watson Test above 1.6, increase in T Test for the observed value, increase in F change, minimal standard error of the estimate, strong midpoint values in letter value display and well behave scatter plots

Keywords: Correction; Detention; Model Violations, Residuals, Linear Regression.

1.0 Introduction

Regression analysis is an important tool of statistical analysis whereby one or more predictors are used to predict a single dependent variable(Y). Regression is used for other things beside prediction, e.g., to find a model. Before the fitted model is used for prediction or other purpose, analysis of residuals is done to check goodness of fits. This study will however examine the uses of residuals in detecting

(i) **Heteroscedasticity**: This occurs when the unequal variance dispersion of the residuals has a horn-shaped pattern

(ii) **lack of Fit**: This arises as a result of insufficient general design matrix ,that is when the residuals follow a systematic pattern(other than a horn shape)

(iii) **Outliers:** These are data points that are far from the other data points being analyzed. Outliers are data points that are far away from the mean. There are rules and conditions to checkmate outliers and to determine whether they should be remove from the regression model or retained for further analyses.

(iv) **Serial Correlation (non-independence):** The residuals are randomly or uncorrelated with time. This assumption is most likely to be violated when data are collected over sequential periods of time series. Autocorrelation exist when the residuals follow oscillatory or cyclical pattern.

All these analysis shows that the linear regression assumption has been violated and the fitted model not good for prediction and the forecast, confidence intervals as well as economic insights yielded by the regression model may be (at best) inefficient or (at worst) seriously biased or misleading. In addition, the results may not be trustworthy, resulting in a Type 1 or Type 11 error, or over or underestimation of significance or effects size. Pedhazur [1] provided an understanding of the situations when violations of assumptions lead to biasness in the regression coefficients. To improve on the model, a transformation is done. Transformation of real and simulated data arises when the residuals do not satisfy the validity assumption of the regression modeling. However, picking a transformation is often a matter of trial and error. Different transformations are tried until one is found for which the residuals seem reasonable [2]. Ostrom [3] outlined four principal assumptions which justify the use of linear regression models for purpose of prediction.

Corresponding author: ***Osemeke Reuben. F.,*** E-mail: reubitosmek@yahoo.com, Tel. +234 8065917356

(i) Linearity of the error term (ii) Independence of the errors (iii)Equal variance dispersion (iv) Normality of the error term

2.0 Theoretical Framework of Residual Violation

Violation of Homoscedasticity:

Heteroscedasticity refers to unequal variance dispersion of the error term over a range of predictor variable. It is visually revealed by a "funnel shape" or horn shape" or "bow shaped" in the plot of the standardized residuals against the estimates ($\hat{y}_i$) or single predictor [4]. The effects of heteroscedasticity is that it does bias coefficient estimates, standard errors of the estimates are incorrect (often underestimates), and the statistical inference are invalid. Some researchers [4, 5] proposed that slight heteroscedasticity has little effect on significance test. When heteroscedasticity is marked, it can lead to serious distortion of findings and seriously weaken the analysis, thus increasing the possibility of Type I error.

Violations of Linearity:

A curve or quadratic trends in the residuals plot indicates curvilinear effects in the error term. Any systematic pattern in the residuals (other than the horn shaped) can indicate lack of fit. Most commonly, one looks for linear or quadratic trends in the residuals. Such trends indicate the existence of effects that have not been removed from the residuals, i.e., effects that have not been accounted for in the model [6]. Authors such as [1], [7] and [4] suggest three primary ways to detect nonlinearity.

(i) The first method is to use the theory of previous research to inform current analysis.

(ii)A preferable method of detention is to examine residual plots of standardized residuals as a function of standardized predicted values.

(iii) The third method of detecting curvilinear is to routinely run regression analyses that incorporate curvilinear components (Square .cubic term; see [8])

Detecting Non-normality of the Error Term Using Letter Value Diagnostics

Letter values are similar to percentiles of the data and are defined by their depth. We use n for the number of observations and [X] for the greatest integer less than or equal to X

Depth of median $\mathrm{D(M)} = \frac{(n+1)}{2}$, Depth of the hinges $\mathrm{d(H)} = \frac{([d(M)]+1)}{2}$

Depth of eights $\mathrm{D(E)} = \frac{([d(H)]+1)}{2}$: Depth of the sixteenths $\mathrm{d(D)} = \frac{([d(E)]+1)}{2}$

To find the letter values, first order the residuals. The lower hinge is the observation at a distance $\mathrm{d(H)}$ from smallest observation; the upper hinge is the observation at a distance $\mathrm{d(H)}$ from the largest observation. Similarly, the lower and upper eights are the observations at a depth $\mathrm{d(E)}$ and so on. The midpoint for a given depth is the average of the upper and lower letter values at that depth. The spread is (upper-lower). If the mid-summaries become progressively larger, the batch is skewed towards the high side. If they decrease steadily, the batch is skewed towards the low side. All this is an indication that the residuals are not normally distributed and a transformation is needed to improve on the model.

Detecting Violations of independence Using Durbin Watson Statistic

Another test for assessing autocorrelation is the use of Durbin-Watson statistic. As a rule of thumb, Durbin Watson should be between 1.5 and 1.6 and above to indicate independence of the observations. Above 3.00 shows a strong independent and any value less than 1.5 lead us to suspect autocorrelation, which is an indication that one of the assumption has been violated [9]

Conditions for assessing outliers in a regression model:

The detecting of outliers is group into two categories

(i) Identifying outliers using graphical method

As a rule of thumb, we can flag any point that is located further than 2*standard deviation above or below the best fit line as an outlier. The standard deviation is computed from the residuals

We can do this virtually in the scatter plot by drawing an extra pair of lines that are 2*SD above or below the best fit line. Any data point that lies outside this extra pair of lines are flagged as potential outliers

(ii**) Identifying outliers using numerical test**

An outlier is detected when the generated residuals are > (2 * SD) above or below the best line fit. The outliers should be removed and re-defined the model again. The value for correlation coefficient must lies between –1 or +1

3.1 Data Analysis to Illustrate Model Violations/Corrections

Detecting Nonlinearity (Curvilinear effects)

Example 1: Simulated data from Eleven observations with X = Ages in days and Y = Dry Weight

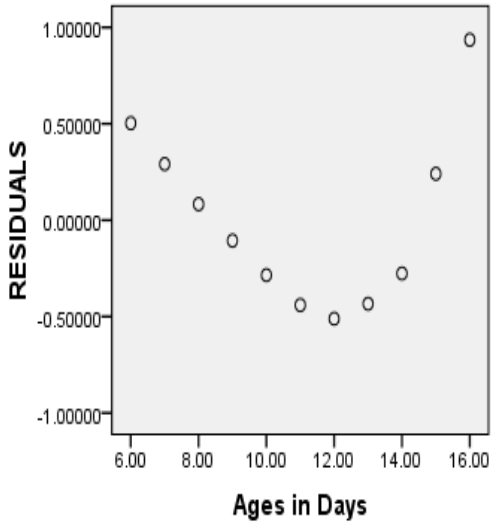


Fig 1: Residual plot versus Ages in Days (Curvilinear Effects)

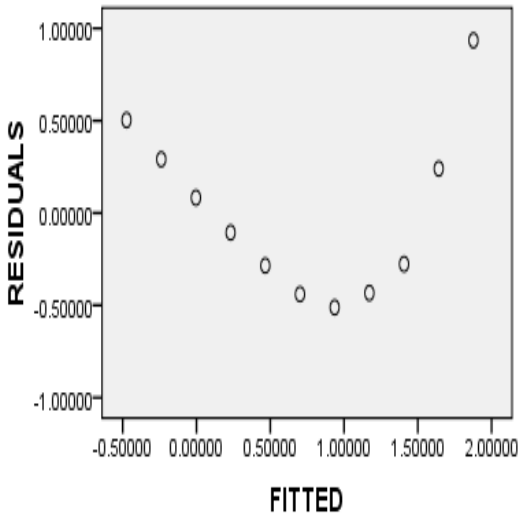


Fig 2: Residual plot against Fitted (Curvilinear Effects)

The scatter plot of Fig 1 and 2 shows a curvilinear trend of the residuals .This is an indication that the linearity of the error term has been violated. The violation shows the fitted model as $\hat{Y}i = -1.885 + 0.235X_i$. The model is inappropriate and not good for prediction .R^2 is 0.744. This represents 74.4% of the Dry Weight as explained by the variation in the Ages in Days(X). The standard error is 0.48185, very minimal

Example 2: Taking a log transformation of Dry Weights(y)

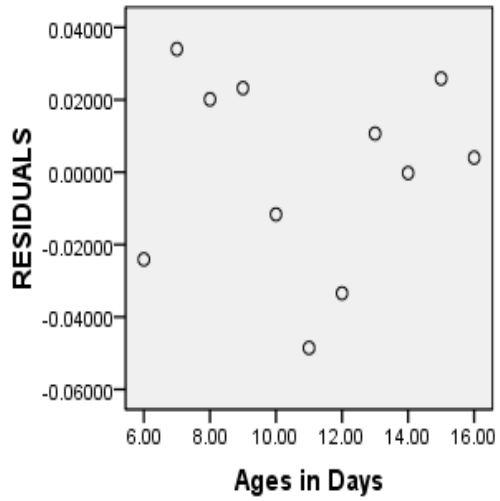


FIG 3: Residuals versus Ages in Days(X). (Randomly Distributed)

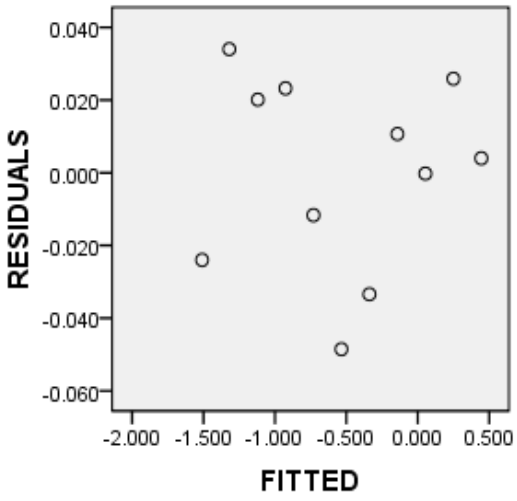


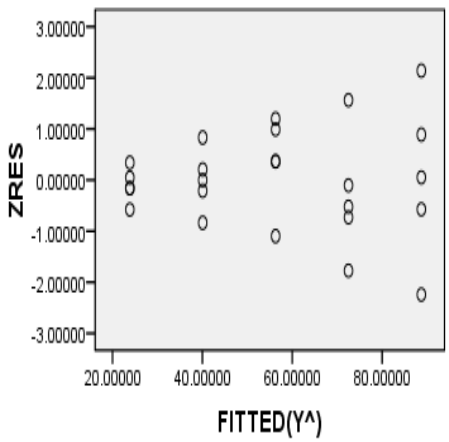
Fig4: Residuals versus Fitted (Randomly Distributed)

With log transformation, R^2 has improved to 99.8%. Standard error has reduced to 0.02807 extremely minimal. There is an improved in the fitted model,

$\hat{Y}i = -2.689 + 0.196X_i$. The residuals scatter plot of Fig3 and Fig4 are evenly distributed which shows that the errors are normally distributed.

Detection of Heteroscedasticity

Example 3: Simulated data with X = Size of Farm and Y = Acres in Corn.(n= 15)



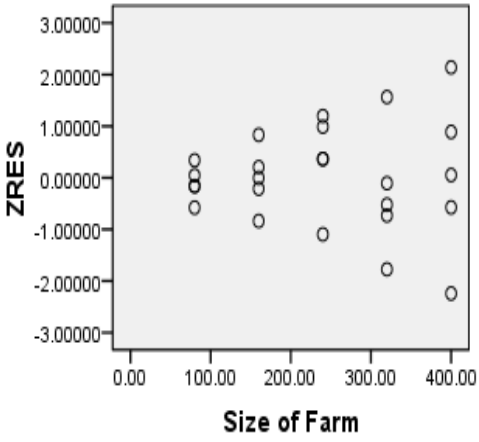


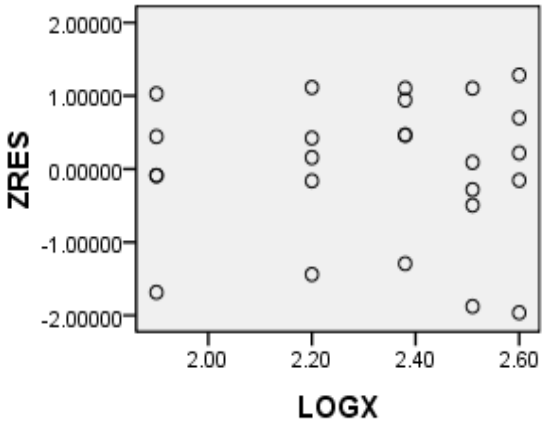
FIG 5: Violations of ZRES plot versus FITTED(Y^). (Heteroscedasticity Detection)

FIG 6: Violation of ZRES plot versus Size of Farm (Heteroscedasticity Detection)

r^2 = 0.499 .This show that there is poor influence of Size of farm of Acres of corn. Standard error of the estimate is 23.89727, a bit high. The model is $\hat{Y}i = 7.620 + 0.203X_i$.

The model is inappropriate. The scatter plot of Fig 5 and 6 display a horn shaped, indicating evidence of heteroscedasticity tendency of the error term.

Example 4: Transformation of LOG *X* and LOG *Y* was used to remove heteroscesdasticity effects



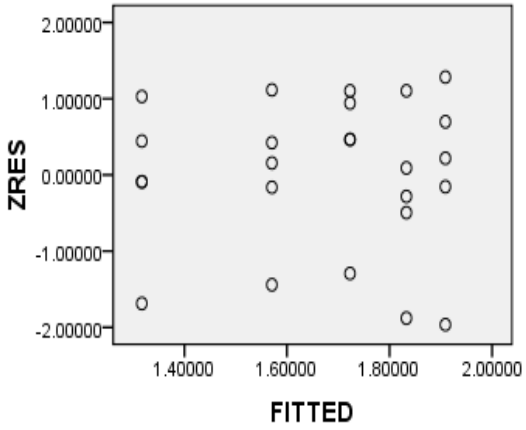


FIG 7 : Validation Effects of ZRES against LOGX :

FIG 8 : Validation Effects of ZRES versus FITTED

R^2 went up to 0.577 very encouraging influence of log transformation of size of farm on Acres in corn. The mode $\hat{y}_i = -0.291 + 0.846X_i$ has improved upon. The standard error for estimate is 0.18788. Durbin Watson = 2.375. P value = 0.0000. The scatter plot of Fig 7 and 8 are evenly distributed and satisfies the regression modeling of homoscedasticity.

Detecting Non-normality of ordered standardized residuals using the Letter Value Display

Table 1: Letter Value Display

	Lower	**Upper**	**Mid**	**Spread**
M = 13	**-0.00250**		**-0.00250**	**-0.00250**
H = 7	**-0.57252**	**0.36387**	**-0.1043**	**0.9364**
E = 4	**-0.83708**	**0.98981**	**0.0764**	**1.8269**
D = 2.25	**-0.93505**	**1.38164**	**0.2234**	**2.3167**
1	**-2.24167**	**2.13985**	**-0.0509**	**4.38152**

We can observe in Table 1 that the mid summary increases gradually, indicating a slight skewness towards the high side. In addition, our spread increase gradually, this indicates non normality of the error term

Transformation to correct non normality of the error term

Many trial by error transformation were done on the variable y and x, but the best accepted transformation was the log transformation of the variable y and x

Table 2: Transformed Letter Value Display of Table 1

	Lower	Upper	Mid	Spread
M = 13	**0.15660**		**0.15660**	
H = 7	**-0.28124**	**0.69736**	**-0.20806**	**0.9786**
E = 4	**-1.44013**	**1.10256**	**-0.168785**	**2.54269**
D = 2.25	**-1.782005**	**1.108685**	**-0.33666**	**2.89069**
1	**-1.96396**	**1.28285**	**-0.340555**	**3.24689**

The mid summary values are almost very approximate. By approximation, the spread are the same across, but 0.9786 is outside the range A little strong relationship .This form of transformation is accepted because the data are very close. The spread shows little upward trend but the trend is much weaker than in the raw data

Detecting auto-correlated effects

Regression Statistics for Package Delivery Store of Fifteen observations with Sales as Y and Customers as X and Week as i, i = 1……15

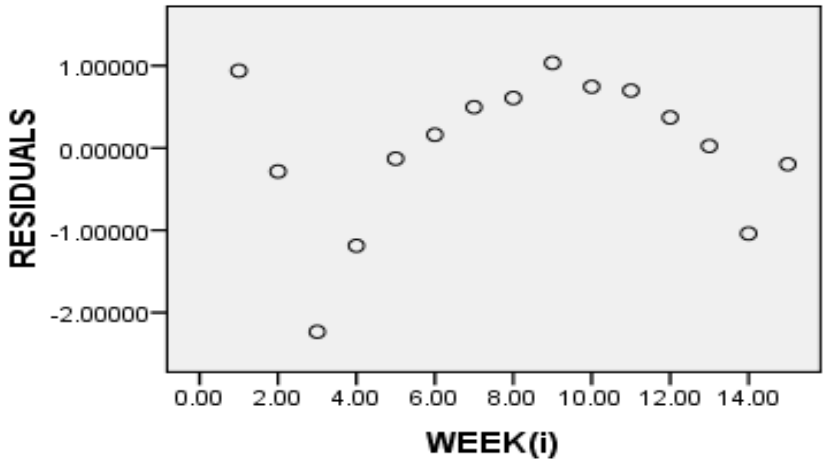


FIG 9: Cyclical Pattern of Residuals plot against Week (i = 1, 2…..15)

R^2 is 0.657. Our Durbin Watson statistic = 0.883. This is low and indicates serious autocorrelation. From the scatter plot of Fig 9, we observe that the point tends to fluctuate up and down in a cyclical pattern. This cyclical pattern gives us strong cause for concern about the autocorrelation of the residuals and, our fitted model

$\hat{Y}i = -16.032 + 0.3076X_i$ is inappropriate because of the presence of serious auto correlation among the residuals. Hence a transformation of the data is needed to correct the auto correlated effects. Our P value for predictor is 0.000245 and standard error = 0.93604.

Example 8: The Log transformation of Customers and Lin of Sales was used to remove serious auto correlated effects

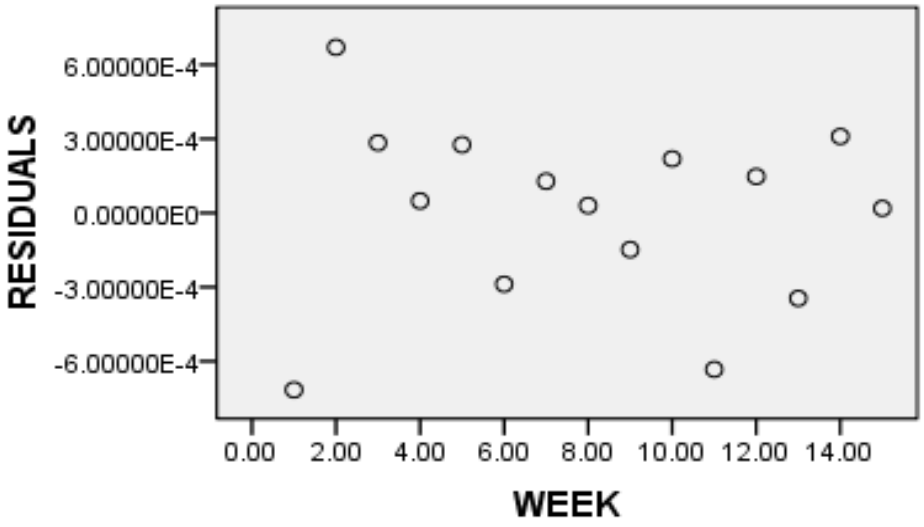


FIG 10: Validation effects of Residual Plot versus Week (i = 1, 2……, 15)

The Log transformation of sales and Lin transformation of Customers was used to remove the auto correlated effects. R^2 went up to 1.000. Durbin Watson went up to 2.559 very high and indication of independence of the error term. Standard error is 0.00039 very minimal. The fitted linear model is $\log\hat{Y}i = -0.009 + 0.435 lin X_i$ .The scatter plots of fig 10 follow homocesdasticity pattern

Assessing Outliers

Simulated data from Eleven observations with X = Third score and omitted Y = Final score were used to assess outlying effects in a regression model

Table 3: Viewing and Assessing Outliers

X	**65**	**67**	**71**	**71**	**66**	**75**	**67**	**70**	**71**	**69**	**69**	
Residual	**35**	**-17**	**16**	**-6**	**-19**	**9**	**3**	**-1**	**-10**	**-9**	**-1**	**SD = 16.4**

Source: Hildebrand & Lyman, 1991)

An outliers is remove in a linear regression when the error (e) > 2*SD, and < than 2*SD where S is computed from the residuals. The standard deviation of the residuals (SD) = 16.4. 2SD = 2(16.4) = 32.8 or less than -32.8.We are looking for all data points for which the residuals greater than 32.8 and less than -32.8 that should be remove from the model. Compare these values to residuals in row 2 of the table 3 and you observed that the highest value for residuals is 35 which is > 2*SD. Therefore, we identified the point (65, 175) as an outlier and should be remove from the model.

Compute a new best fit and correlation coefficient using 10 remaining point. The new best line of fit $\hat{Y}i = -.355.19 + 7.38X_i$. The new line of correlation coefficient

r = 0.9121, r^2 = 0.832, standard error = 9.275, P value = 0.000, n = 10 and Std for the residuals = 8 is a strong correlation than the original data with regression statistics of r = 0.6631, r^2 = 0.440, std error = 16.4124, high p value of 0.026 and fitted model of

$\hat{Y}i = -173.513 + 4.83X_i$. This is really a bad regression due to outlying effects.

Finally r = 0.9121 is closer to 1 which satisfied the condition of correlation coefficient of 1 or -1.The regression is good and this means that the new line is a better fit to the 10 remaining data values. The line can better predict the final exam score given the third exam score. Further analysis shows that is no outlier detected because 2*s = 16 and -16 and none of them is greater than 16 and less than -16, so there are no further outlier

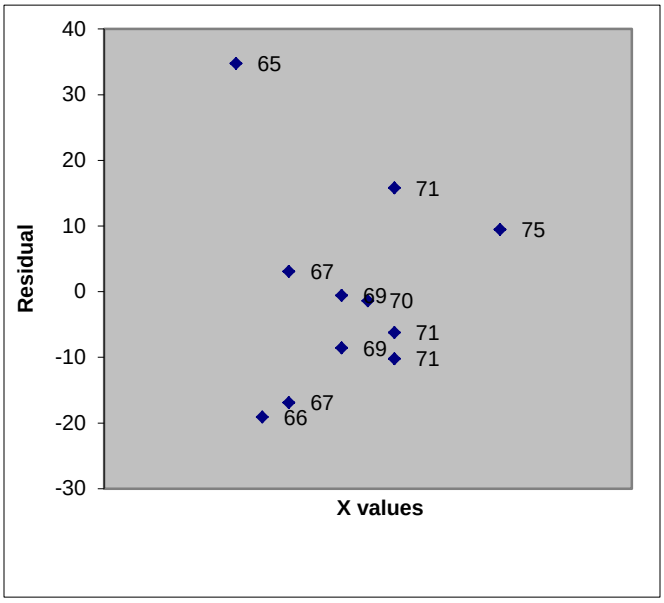


FIG 11: Scatter plot of full Regression
(Residuals versus X values)

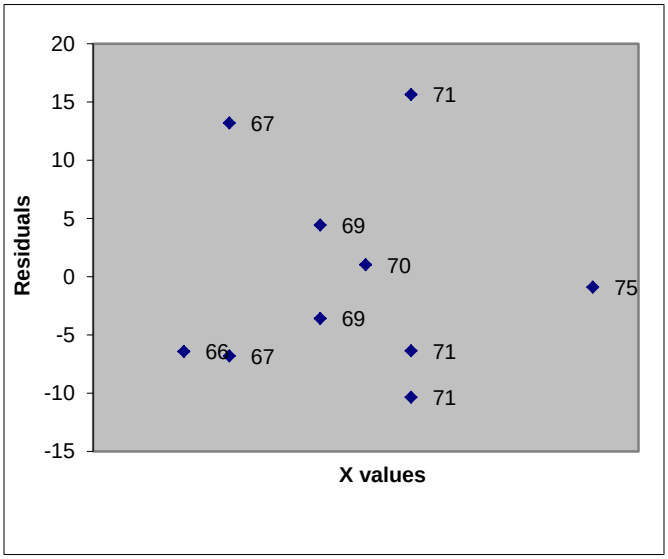


FIG 12: Removing Row 1 from the model
(Residuals versus X values)

3.0. Conclusion

The goal of this article was to raise awareness of the importance of checking model assumptions when they are violated due to heteroscedasticity, curvilinear effects, outliers, non-normality, non-independence and clustering of data points of the error term assumptions

The potential model violation was examined through examination of residual plots against the predictors and numerical statistics.

Using these techniques, the researcher was able to have an insight in the violations of regression assumption. The values of employing these techniques are well documented in books [6, 9 - 11]. Error violence were corrected using plausible trial by error transformation and after transformation, the model were improved upon.
In all these, analysis of residuals after the correction , shows that the correction were improved upon as shown in multiple r, coefficient of determination(r^2),well behave plots, decrease in standard errors and increase in Durbin Watson statistic.
We therefore regard residual analysis as an indispensable tool of regression analysis. It facilitates the job of the analysis

4.0. Recommendation

In dealing with the uses of residuals in detecting invalidity of the assumption, certain measures have to be put into consideration to ensure a very easy analysis of residuals
In practice, the real question is not whether the data are non normal, but whether they are sufficiently non-normal to invalidate a normal approximation. This is a more difficult question to address
The transformation of log to base 10 functions is most frequent used to correct model violations. This is because, for a linear model to base 10 functions, the additive effects of the predictor variables transform multiplicative effects on the original scale. If multiplicative effects seem reasonable, the log transformation may be appropriate
Moreover, methods for detecting non-normality are often sensitive to inequality of variance, so the use of residuals can make it appear that the error are not normal even when they are normal.
Cook and Weisberg [6] proposed that in detecting heteroscedasticity of the variance of the residuals, it is important to standardize the residuals, because ordinary residuals have heteroscedasticity tendency, so before they are used in checking for equality of the variance of the observations, they need to be standardized

REFERENCES

[1] Pedhazur, E.J., (1997). Multiple Regressions in Behavioral Research (3rd Ed). Orlando, FL: Harcourt Brace

[2] Atkinson, A. C. (1985). Plots, Transformations and Regression, Oxford, U.K: Oxford Publications

[3] Ostrom, C. W., Jr. (1990). Time Series Analysis Regression Techniques, Second Edition: Quantitative Application in the social Sciences, v.06-009: Newbury Park, Sage Publications.

[4] Berry, W.D. & Feldman's. (1985). Multiple Regression in Practice. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no.07-050).Newbury Park, CA: Sage

[5] Tabachnick, B.G. Fidell, L .S. (1996). Using Multivariate Statistics (3rd Ed) .New York: Harper Collins College Publishers.

[6] Cook, R.D. & Weisberg, S. (1982). Residual and Influence in Regression, New York: Chapman and Hall

[7] Cohen, .J. & Cohen, P. (1983). Applied Multiple Regression & Correlation analysis for the behavioral sciences, 2ed. Hillsdale, N.J: Lawrence Erlbaum Associates, Inc.

[8] Quandt,G.,(1976).Nonlinear Methods in Econometrics.North Holland Publisher 2ed Edition

[9] Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics", New York: John Wiley.

[10] Gunt, R.F. & Mason, R.L. (1980), "Regression Analysis and its Application", New York: Marcel Dekker

[11] Montgomery, D.C., & Peck, E.A. (1982), "Introduction to linear Regression Analysis", New York: John Wiley