AN ANALYSIS OF RESIDUALS IN MULTIPLE REGRESSIONS

Ahmad A. Suleiman[1], Usman A. Abdullahi[2], Umar A. Ahmad[3]

[1, 2, 3]Postgraduate Student, Department of Mathematic / Statistics, Sharda University, Greater Noida, (India)

ABSTRACT

This paper concentrates on residuals analysis to check the assumptions for a multiple linear regression model by using graphical method. Specifically, we plot the residuals and standardized residuals given by model against predicted values of the dependent variables, normal probability plot, histogram of residuals and Quantile plot of residuals. Finally, we explained the concept of heteroscedasticity which we used to check the assumption that the residuals in regression model the same variance. As an example, a formal method to detect the presence of heteroscedasticity by Breusch Pagan method using eview was presented.

I. INTRODUCTION

The main aim of regression modelling and analysis is to develop a good predictive relationship between the dependent (response) and independent (predictor) variables. Regression diagnostics plays a vital role in finding and validating such a relationship. In this study, we discuss issues that arise in the development of a multiple linear regression model. Consider the following standard multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

where Y is a response variable and $X's$ are predictor variables, $\beta's$ are the (regression) parameters to be estimated from data, and ε is the error or residual.

The validity of the inference methods depends on the error term ε, satisfying these assumptions;

- **Independence:** Observations (and hence residuals) are statistically independently distributed.
- **Normality:** The residuals are normally distributed with zero mean.
- **Homoscedastiticity:** All the observations (and hence residuals) have the same variance.
- **Multicollinearity:** No linear correlation between independent variables

II. METHOD AND ANALYSIS

Here is a hypothetical data on Consumption, Export and GDP

CONSUMPTION	EXPORT	GDP
50.35718	1.436314	35.06
50.44603	1.414639	35.66
57.87973	1.529996	37.83
72.30876	1.746588	41.4

77.65894	1.801414	43.11
80.01789	1.808723	44.24
103.396	2.092189	49.42
111.4546	2.158299	51.64
126.315	2.292468	55.1
138.3544	2.384188	58.03
149.9102	2.462792	60.87
164.6521	2.56228	64.26
187.6525	2.71842	69.03
195.7883	2.746364	71.29
214.2388	2.846271	75.27
241.5957	2.994864	80.67
288.8777	3.24181	89.11
301.7072	3.274088	92.15
303.5576	3.245564	93.53
292.6464	3.148417	92.95
271.2281	2.991046	90.68
291.739	3.068353	95.08
305.7957	3.105471	98.47
329.3367	3.186307	103.36
366.2961	3.320908	110.3
388.546	3.379249	114.98
433.4515	3.52285	123.04
503.9317	3.740863	134.71
513.4575	3.732336	137.57
505.2289	3.663468	137.91
583.8524	3.874784	150.68
582.0886	3.827768	152.07
635.0129	3.940019	161.17
716.5901	4.115023	174.14
634.7767	3.855544	164.64
706.7427	4.004662	176.48

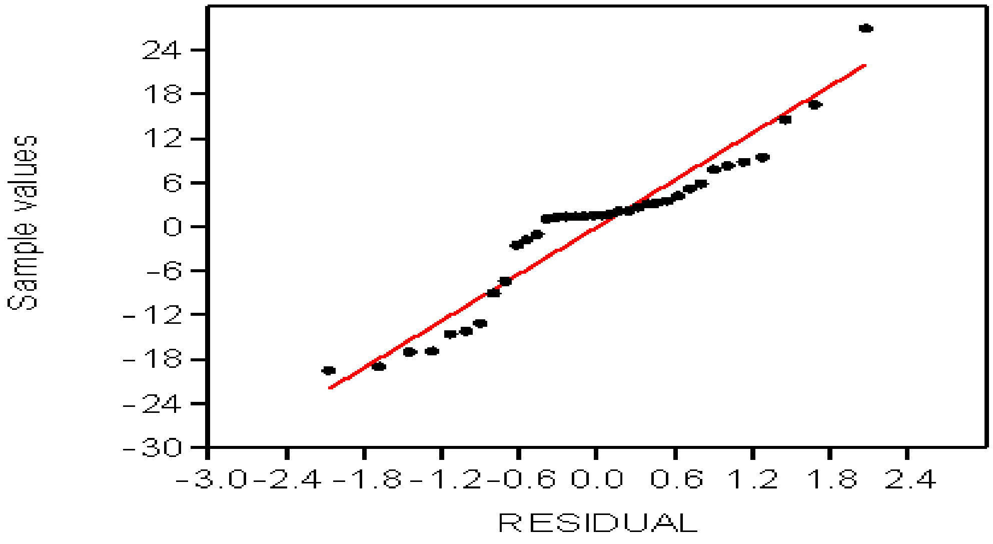
2.1 Regression Diagnostics

Saving Residuals for Diagnosis:

There are many diagnostics we can perform on the residuals. Here are the most important ones:

- **Normal Probability Plot**

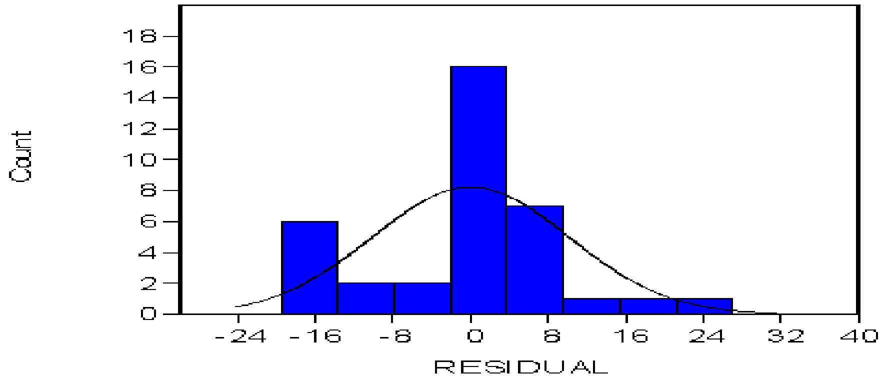
To diagnose if the errors are normally distributed, we draw a normal probability plot of the residuals. The residuals should fall approximately on a diagonal straight line in this plot,



Normal Probability Plot

- **histogram of the Residuals**

We can also plot a histogram to see if they are lumpy in the middle symmetric tails.

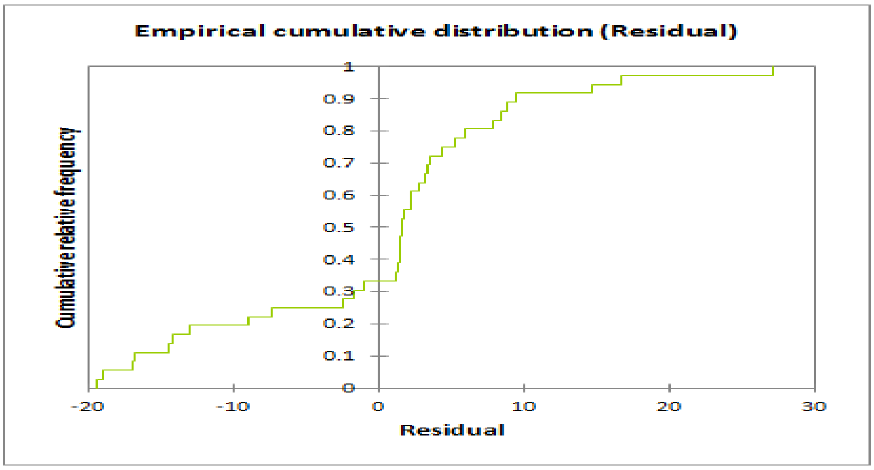


Histogram of Residuals

From our histogram, the residuals appear to be normally distributed.

- **Quantile Plot of Residuals**

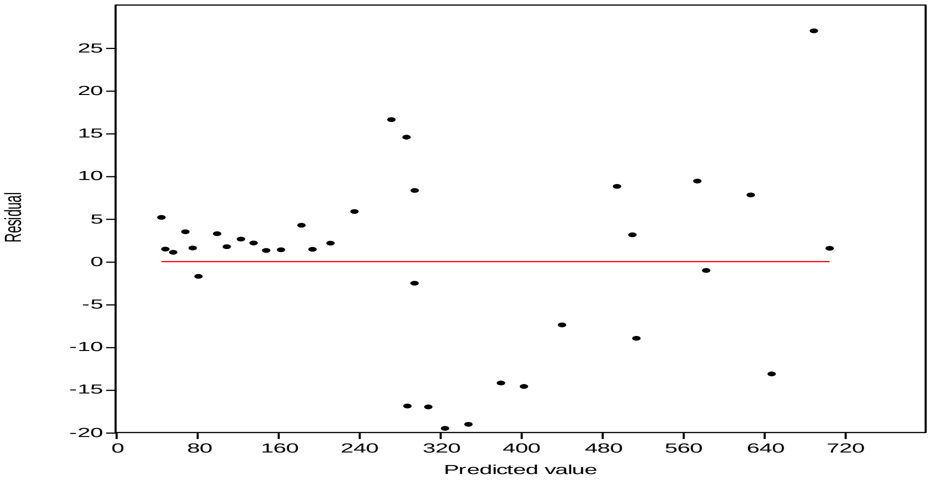
We can make a Quantile Plot where the residuals are plotted against their percentage (empirical cumulative probability) point (0 to 1): if normality holds shis should have an S-shape.



Quantile Plot of Residuals

The S-shape of the curve seems to suggest that normality assumption is satisfied.

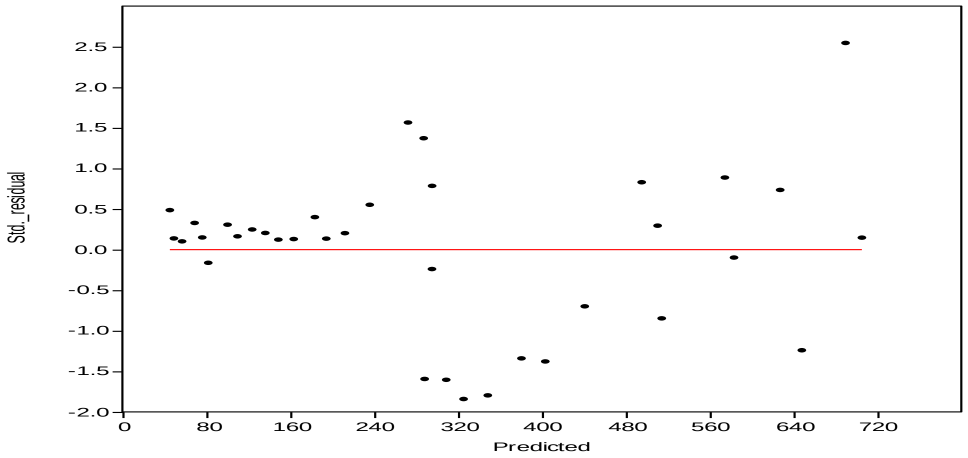
To examine if the errors are independent. We look at the plot of residuals against estimated values to ensure that the residuals are randombly scattered above and below the 0 horizontal.



Although the mean of the residual may be accepted to be zero at each x-value, the variance seems to increase with x-value suggesting a possible violation of thehomoscedasticity assumption.

- **Standardized Residuals**

The Standardized residuals is more useful since they are more easy to interpret. We plot Standardized residuals against estimated values to identified outliers in the dependent variable space. Large values (greater than 2 or 3 in absolute magnitude) indicate posssible problems.



Case no. 34 with a Standardized value of 2.246 can be considered an outlier. If we delete this case from the data set and recompute the regression, the fit become better.

2.2 Removal of Heteroscedasticity from Regression Mode

We want remove heteroscedasticity from the regression model because heteroscedasticity is not desirable that is the residuals should be homoscedastic.

There are many ways to remove heteroscedasticity from the model. One of the most popular ways is to convert all the variables into log, which is known as log transformation.

Let us consider another multiple regression example with quite a few predictor variables. In our model, we have three variables such as consumption, export and growth domestic product (gdp). Here consumption is the dependent variable while the rest two are independent variables.

If we see heteroscedasticity in the model after estimation, we need to convert all the three variables into log that is: Consumption >log(consumption), export > log(export), gdp> log(gdp).

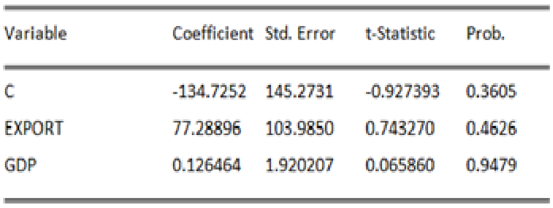
Once we run the model with log variables, heteroscedasticity will be removed and homoscedasticity will be appeared. As we know, homoscedasticity is desirable.

Our hypothesis:

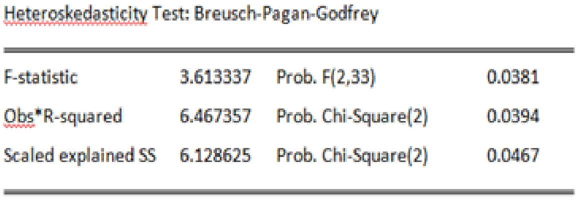
Null hypothesis: Homoscedasticity

Alternative hypothesis: Heteroscedasticity

Basic Software output yield:

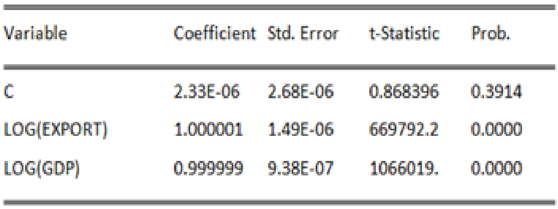


Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-134.7252	145.2731	-0.927393	0.3605
EXPORT	77.28896	103.9850	0.743270	0.4626
GDP	0.126464	1.920207	0.065860	0.9479

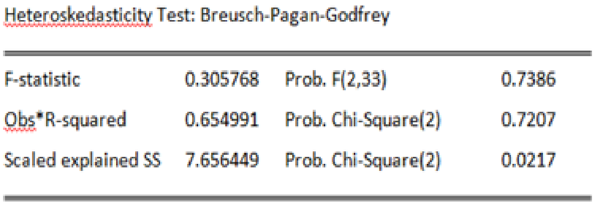


Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	3.613337	Prob. F(2,33)	0.0381
Obs*R-squared	6.467357	Prob. Chi-Square(2)	0.0394
Scaled explained SS	6.128625	Prob. Chi-Square(2)	0.0467



Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.33E-06	2.68E-06	0.868396	0.3914
LOG(EXPORT)	1.000001	1.49E-06	669792.2	0.0000
LOG(GDP)	0.999999	9.38E-07	1066019.	0.0000



Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	0.305768	Prob. F(2,33)	0.7386
Obs*R-squared	0.654991	Prob. Chi-Square(2)	0.7207
Scaled explained SS	7.656449	Prob. Chi-Square(2)	0.0217

Now, we are to check whether this model has heteroscedasticity or not. Using Breusch- Pagan- Godfrey test we have p-value corresponding to R^2 is 0.0394 which is less than 5%, hence we reject the null hypothesis and accept the alternative hypothesis. Meaning that this model got heteroscedasticity. In other words, the residuals are heteroscedasticity so that the model is not desirable.

After transforming the variables Breusch-Pagan-Godfrey shows that the p -value corresponding to the observed R^2 is 0.7207, which is more than 5%. Meaning that, we cannot reject the null hypothesis. Hence the model is homoscedasticity which is desirable.

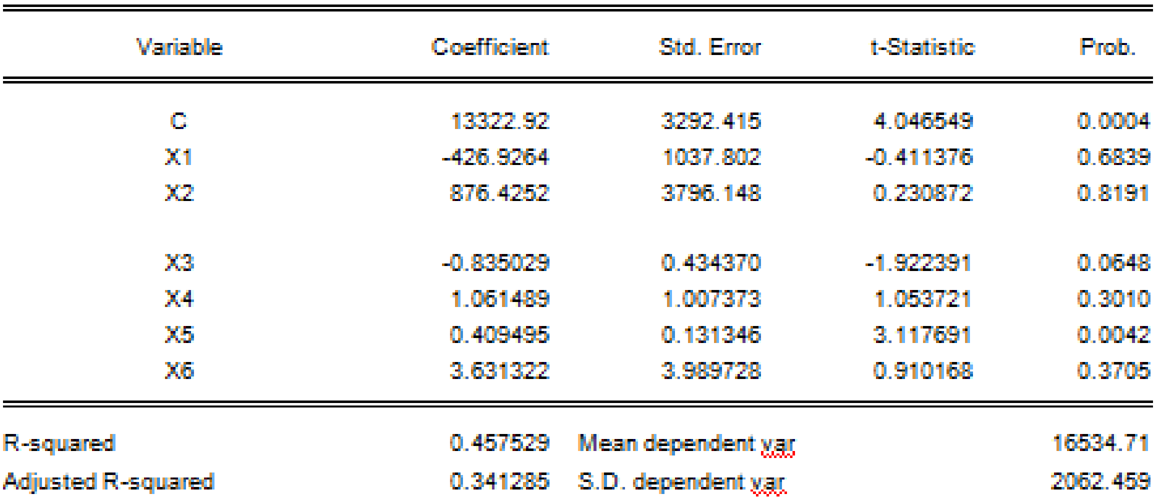
2.3 Multicollinearity Problem and Regression Model

How multicollinearity affects any estimated regression model?

Here is our model:

$$Y = C + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 \text{ ....} (1.1)$$

Here Y is the dependent variable and the rest are independent variables. After estimating Model (1.1), we saw that only X_5 is significant while others are not. We suspect that, there is a problem of multicollinearioty in Model (1.1) that is why most of the variables have become insignificant.



Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13322.92	3292.415	4.046549	0.0004
X1	-426.9264	1037.802	-0.411376	0.6839
X2	876.4252	3796.148	0.230872	0.8191
X3	-0.835029	0.434370	-1.922391	0.0648
X4	1.061489	1.007373	1.053721	0.3010
X5	0.409495	0.131346	3.117691	0.0042
X6	3.631322	3.989728	0.910168	0.3705
R-squared	0.457529	Mean dependent var		16534.71
Adjusted R-squared	0.341285	S.D. dependent var		2062.459

III. CONCLUSION

1. We estimate the value of the regression residuals for each value of y :

$$\hat{\varepsilon} = y - \hat{y}$$

Whichis the observed value $-$ the predicted $($or expected$)$ value.

2. We made sure the removal of multicollinearity by dropping the appropriate highly correlated independent variables before studying the residuals.
3. We have dealt with an example dealing with GDP in chapter four where we detected and removed heteroscedasticity by the methods suggested by Bruesch-Pagan.

Another problem that can affect adversely our study of multiple regression is heteroscedasticity.

4. In conclusion we have dealt with

- estimates of regression coefficients;
- their standard errors, confidence intervals and tests of their significance;
- analysis of variance for regression which produces an overall test of significance and an estimate of the error variance as the residual (error) mean-square;
- multiple correlation coefficient, its square, and its adjusted value, which give a measure of how much of the variation has been captured by the predictor variables and hence how useful the regression is;
- Using the saved residuals we can
  - ➢ make suitable plots to examine the assumption of normality;
  - ➢ carry out a formal test of significance for normality;
  - ➢ make suitable plots to examine the assumption of homoscedasticity;
  - ➢ make suitable plots to examine the assumption of independence.

Thus some basic diagnosis can be performed of the validity of assumptions under which a standard regression analysis is carried out using this regression output.

IV. ACKNOWLEDGEMENTS

Our infinite gratitude goes first and foremost to Almighty Allah for sparing my life till this moment.

Many special thanks and appreciation goes to our honorable supervisor and at the same time our esteemed lecturer **Prof. U.V. Balakrishnan** who his suggestions, corrections and encouragements made this work reality. We are also grateful to the entire lecturers in the Department of Statistics in person of **Dr. N.M. Chahda, Dr. SwetaSrivastav, Dr. Krushidalam**and others whose wealth of experience and knowledge I have benefited.

REFERENCES

[1]. AshishSen and Muni Srivastava, Regression Analysis: Theory, Methods, and Applications,Springer-Verlag, New York, 1990, p. 92. Notation changed.

[2]. Belsley, David A.; Kuh, Edwin; Welsch, Roy E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley. ISBN 0-471-05856-4.

[3]. C. R. Rao, Linear Statistical Inference and Its Applications, John Wiley & Sons, New York, 1965, p. 258.

[4]. D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.

[5]. D. E. Farrar and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem

[6]. Revisited," Review of Economics and Statistics, vol. 49, 1967, pp. 92–107.

[7]. Douglas Montgomery and Elizabeth Peck, Introduction to Linear Regression Analysis

[8]. H. Glejser, "A New Test for Heteroscedasticity,'' Journal of the American Statistical Association, vol. 64, 1969, pp. 316–323.

[9]. Hill, R. Carter; Adkins, Lee C. (2001). "Collinearity". In Baltagi, Badi H. A Companion to Theoretical Econometrics. Blackwell. pp. 256–278. doi:10.1002/9780470996249.ch13. ISBN 0-631-21254-X.

[10]. J. T. Webster,"Regression Analysis and Problems of Multicollinearity," Communications in Statistics A, vol. 4, no. 3, 1975, pp. 277–292; R. F. Gunst.

[11]. Johnston, John (1972). Econometric Methods (Second ed.). New York: McGraw-Hill. pp. 159–168.

[12]. John Wiley & Sons, New York, 1982, pp. 289–290. See also R. L. Mason, R. F. Gunst.

[13]. Kmenta, Jan (1986). Elements of Econometrics (Second ed.). New York: Macmillan. pp. 430–442. ISBN 0-02-365070-2.

[14]. Maddala, G. S.; Lahiri, Kajal (2009). Introduction to Econometrics (Fourth ed.). Chichester: Wiley. pp. 279–312. ISBN 978-0-470-01512-4.

[15]. R. Koenker, "A Note on Studentizing a Test for Heteroscedasticity," Journal of Econometrics, vol. 17, 1981, pp. 1180–1200.

[16]. R. L. Mason, "Advantages of Examining Multicollinearities in Regression Analysis," Biometrics, vol. 33, 1977, pp. 249–260.

[17]. T. Breusch and A. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient

[18]. Variation,'' Econometrica, vol. 47, 1979, pp. 1287–1294.