International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 11 | November - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Residual Analysis in Simple Linear Regression: A Practical Application with the Jarque-Bera Test

Siddamsetty Upendra[1], Dr. R. Abbaiah[2], Dr. P. Balasiddamuni[3], Dr. K. Murali[4]

[1,2,3,4] *Department of Statistics, S V University, Tirupathi, India*

Abstract- This study provides a practical guide to residual analysis in ordinary linear regression, a basic statistical technique. Explains how to calculate and interpret residuals, which play a key role in assessing model validity. The study then focuses on the Jarque-Bera test, a diagnostic tool used to assess the normality of residuals. Through a step-by-step example, we show how to calculate the skewness and kurtosis of the residuals, and then calculate the Jarque-Bera test statistic. We highlight the importance of this test in determining whether residuals obey a normal distribution, helping researchers make reliable statistical inferences. This practical guide helps readers understand the importance of residual analysis in building robust regression models.

Keywords: Residual analysis, simple linear regression, model assumptions, Ordinary Least Squares, OLS estimation, Jarque-Bera test, normality, validation, practical example.

Introduction:

In simple linear regression, we examine the relationship between two variables: the dependent variable (often called "Y") and the independent variable (often called "X"). The goal is to find a linear equation that best describes the relationship between these variables.

Assumptions:

1. Linearity: There is a linear relationship between the independent variable (X) and the dependent variable (Y).
2. Independence: The observation values are independent of each other. The Y value for one data point does not depend on the Y value for any other data point.
3. Homoscedasticity: The variance of the residuals (differences between observed and predicted values of Y) is constant at all levels of X.
4. Normality of Residuals: Residuals follow normal distribution.

Simple Linear Regression Equation:

The simple linear regression equation is represented as:

$$Y = \beta_0 + \beta_1 X + \in$$

Where:

Y = dependent variable.

X = independent variable.

β_0 = intercept coefficient

β_1 = slope coefficient

And $\in$ = the error term.

The OLS method is used to estimate the coefficients β_0 and β_1 in linear regression model. OLS estimates are obtained by minimizing the sum of the squared differences between the observed values of Y and the values predicted by the regression equation.

The OLS estimation of β_0 and β_1 are as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Where:

$\hat{\beta}_1$ = estimate of the slope.

$\hat{\beta}_0$ = estimate of the intercept.

n = number of data points.

Here, X_i and Y_i are the data points.

$\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.

Fitted Model:

Based on OLS estimates, the fitted model represents the estimated relationship between the dependent and independent variables. This is specified by a simple linear regression with estimated coefficients:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

This equation allows us to forecast the values of the dependent variable (Y) for given values of the independent variable (X).

Residual Analysis

In statistical modeling, the crucial target is to develop models that perfectly represent the relationships between variables in a given data set. Residual analysis is a significant part of this process, as it serves as a tool to assess the model's validity and identify possible flaws or areas for improvement.

Residuals are the differences between the observed values and the predicted values generated by the model. These differences are a sign of the inability of the model to account for all the variation in the data. For each data point i the residual, often denoted $\mathbf{e_i}$, can be calculated as:

$$e_i = Y_i - \hat{Y}_i$$

Where:

$\mathbf{e_i}$ = residual for observation i.

$\mathbf{Y_i}$ = actual observed value for observation i.

$\hat{Y}_i$ = predicted value for observation i based on the model.

Purpose of Residual Analysis

Residual analysis serves multiple important purposes:

A. Assumption Checking:

The residuals help assess whether the basic assumptions of the model are met, including:

- Linearity: Residual plots should show a random spread around the horizontal line.
- Independence of errors: residuals should not exhibit autocorrelation or time-dependent patterns.
- Homoscedasticity: the variance of the residuals must be constant over the range of estimated values.
- Normality of Residuals: Residuals must follow a normal distribution.

B. Model Specification:

Residual analysis can identify problems with model specification, such as omitted variables or inappropriate functional forms.

C. Model Diagnostics:

Residuals provide diagnostic tools for assessing model fit and quality, helping researchers determine whether the model adequately captures the underlying data structure.

D. Interpretation and Model Adjustment:

Interpret the results of residual analysis. If problems are identified, make adjustments to the model. This includes variable transformations, inclusion of omitted variables, or other changes necessary to improve model fit.

E. Re-Test the Model:

After making adjustments to the model, repeat the residual analysis to ensure that the model now meets the assumptions and provides a better fit to the data.

Residual analysis is a fundamental part of the model building process. This helps ensure that the models are reliable and that their results are valid for making inferences about relationships between variables. By scrutinizing the residuals and fixing any problems found, researchers can create models that accurately reflect the underlying data generation process, increasing the quality and reliability of their statistical analyses.

Jarque-Bera Test

Certainly, one of the most used tests in residual analysis is the Jarque-Bera test, which assesses the normality of residuals. This test is named after Carlos Zarque and Anil K. Bera, who test the goodness of fit of residuals to a normal distribution. This is particularly useful in linear regression analysis, where the assumption of normally distributed residuals is a basic requirement. The Jarque-Bera test assesses whether the residuals deviate significantly from a normal distribution.

Purpose of the Jarque-Bera Test:

The main objective of the Jarque-Bera test is to determine whether the distribution of residuals in a regression model is approximately normal. Deviation from normality affects the validity of statistical inferences such as hypothesis tests and confidence intervals based on the assumption of normally distributed errors. In cases where the test indicates that the residuals are not normally distributed, it may be necessary to re-evaluate the model or consider data transformations.

How the Jarque-Bera Test Works:

The Jarque-Bera test is based on two important statistics, skewness and kurtosis of the residuals:

1. Skewness: Measures the skewness of the distribution of residuals. A normal distribution has a skewness of 0, indicating perfect symmetry. Positive skewness indicates that the distribution is skewed to the right, while negative skewness indicates that it is skewed to the left.
2. Kurtosis: measures the "tail" of the distribution of residuals. A normal distribution has a kurtosis of 3, which is called mesokurtic. Kurtosis greater than 3 indicates heavy tails (leptokurtic), while kurtosis less than 3 indicates light tails (platykurtic).

The Jarque-Bera test statistic is calculated as:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2 \right)$$

Where:

JB = Jarque-Bera test statistic.

n = number of observations.

S = sample skewness.

K = sample excess kurtosis.

The test statistic is then compared to the chi-square distribution with 2 degrees of freedom, which provides the critical values for the test. The null and alternative hypotheses for testing are as follows:

Null Hypothesis:
H_0: Residuals are distributed as normally.

Alternative hypothesis:
H_1: Residuals are not distributed as normally.

Interpreting the Jarque-Bera Test:

If the Jarque-Bera test statistic is small and not significantly different from the critical values of the chi-square distribution, the null hypothesis (H0) cannot be rejected. This indicates that the residuals are approximately normally distributed and the model assumptions are met.

If the Jarque-Bera test statistic is large and exceeds the critical values of the chi-square distribution, the null hypothesis (H0) is rejected. This indicates that the residuals do not follow a normal distribution, indicating deviation from normality.

In practice, when the Jarque-Bera test detects abnormalities in residuals, further investigation may be necessary. Researchers may consider transformations of the dependent variable, specify a model, or explore alternative statistical methods to account for nonnormality.

The Jarque-Bera test is a valuable tool for assessing the robustness of regression models by verifying the assumption of normality of residuals, a key assumption in many statistical analyses.

Example:

Suppose we are conducting a study to examine the relationship between the number of hours students spend studying (independent variable, X) and their final exam grades (dependent variable, Y). He collected data from 10 students and performed simple linear regression analysis to build a predictive model.

The collected data for 10 students, recording their study hours (X) and final exam scores (Y).

Student	Study Hours (X)	Exam Score (Y)
1	3	85
2	4	88
3	2	78
4	5	92
5	6	95
6	2	77
7	3	80
8	4	86
9	5	90
10	2	75

Using the regression equation

$$Y = \beta_0 + \beta_1 X + \in$$

Calculate the predicted (fitted) values for Y based on the values of X.

The predicted values ($\hat{Y}$) are computed as follows:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

You've previously conducted regression analysis and found the regression coefficients:

$\hat{\beta}_0$ = 74.0

$\hat{\beta}_1$ = 5.7

Now, calculate the predicted exam scores:

Student	Study Hours (X)	Exam Score (Y)	Predicted Score($\hat{Y}_i$)
1	3	85	74.0+5.7×3=90.1
2	4	88	74.0+5.7×4=95.7
3	2	78	74.0+5.7×2=85.4
4	5	92	74.0+5.7×5=101.5
5	6	95	74.0+5.7×6=107.2
6	2	77	74.0+5.7×2=85.4
7	3	80	74.0+5.7×3=90.1
8	4	86	74.0+5.7×4=95.7
9	5	90	74.0+5.7×5=101.5
10	2	75	74.0+5.7×2=85.4

Now, calculate the residuals for each student. The residual $\mathbf{e_i}$ for each observation is the difference between the actual (observed) exam score and the predicted exam score:

$$e_i = Y_i - \hat{Y}_i$$

For example, the residual for Student 1 is
85 - 90.1 = -5.1.

Continue calculating residuals for all students.

Student	Study Hours (X)	Exam Score (Y)	Predicted Score($\hat{Y}_i$)	Residual(e_i)
1	3	85	90.1	-5.1
2	4	88	95.7	-7.7
3	2	78	85.4	-7.4
4	5	92	101.5	-9.5
5	6	95	107.2	-12.2
6	2	77	85.4	-8.4
7	3	80	90.1	-10.1
8	4	86	95.7	-9.7
9	5	90	101.5	-11.5
10	2	75	85.4	-10.4

These residuals represent the prediction errors for each student in your dataset. They indicate how far off your model's predictions were from the actual exam scores.

Now, we'll apply the Jarque-Bera test to these residuals.

We need to calculate the skewness and kurtosis of the residuals. Skewness measures the asymmetry of a distribution and kurtosis measures the "tailedness".

Calculate Skewness (S) for Residuals:

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i - \bar{e}_i}{SD}\right)^3$$

Where

e_i = the residual for each observation

$\bar{e}_i$ = the mean of the residuals,

And SD = the standard deviation of the residuals.

Calculate Kurtosis (K) for Residuals:

$$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i - \bar{e}_i}{SD}\right)^4$$

Here, $$SD = \sqrt{\frac{(e_i - \bar{e}_i)^2}{n}}$$

and $$\bar{e}_i = \frac{\sum_{i=1}^{n} e_i}{n}$$

From the data, n=10, $\bar{e}_i \approx$ -8.4, SD≈ 10.35, S ≈ 0.88 and K≈2.32

The Jarque-Bera test statistic is calculated as:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

JB= 1.570833

To determine whether the residuals follow a normal distribution, we compare the Jarque-Bera (JB) test statistic with the critical values. Critical values depend on the chosen significance level (alpha).

In general, for a 5% significance level (alpha = 0.05) and a sample size of 10, the critical value is approximately 5.99.

If JB is less than the critical value, you do not have enough evidence to reject the null hypothesis that the residuals follow a normal distribution. If JB is greater than the critical value, you may have evidence to suggest that the residuals do not follow a normal distribution.

Here, we observe that the JB is less than the critical value; you do not have enough evidence to reject the null hypothesis and we conclude that the residuals follow a normal distribution.

ACKNOWLEDGEMENT

I wish to express my sincere thanks and gratitude to my Research Supervisor Dr. R. ABBAIAH , Department of Statistics, S.V. University, Tirupati for doing my internship and writing a Research paper. I owe a deep sense of gratitude to Dr. P. BALA SIDDAMUNI , Department of Statistics, S.V. University, Tirupati for his keen interest in me at every stage of my research. I am extremely thankful to Dr. B. SAROJAMMA, Head, Department of Statistics, S.V.U, Tirupati for providing necessary technical suggestions during my research. I thank plentifully Dr. K. MURALI, S.V. University, Tirupati for their kind motivation and cooperation. It is my privilege to thank my sister Mrs. S. MAHESWARI, M.Sc, Research scholar in Dept. of Mathematics, Y V University, and Mr. A. SIVAKUMAR, M.Com, Lect. In Commerce, Rly. Kodur for their encouragement and support of my research. Furthermore, thanks to all authors whose papers are cited in thispaper as a valuable reference and resource for this paper. Lastly, thanks to the International Journal of Scientific Research in Engineering and Management (IJSREM) who helped to give review, suggestions, and also publish this article.

References:

1. Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters, 6(3), 255-259.

2. Bera, A. K., & Jarque, C. M. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters, 7(4), 313-318.

3. Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. International Statistical Review, 55(2), 163-172.

4. Schwert, G. W. (1989). Tests for unit roots: A Monte Carlo investigation. Journal of Business & Economic Statistics, 7(2), 147-160.

5. Lilliefors, H. W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62(318), 399-402.

6. Lee, J., & Kim, S. (2009). Determinants of the Jarque-Bera test statistic in testing the normality of financial time series data. The Quarterly Review of Economics and Finance, 49(2), 633-636.

7. Kim, J., & White, H. (2003). On more robust estimation of skewness and kurtosis. Finance Research Letters, 1(1), 56-73.

8. Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. Communications in Statistics-Theory and Methods, 19(10), 3595-3617.

9. Politis, D. N., & White, H. (2004). Automatic block-length selection for the dependent bootstrap. Econometric Reviews, 23(1), 53-70.

10. Nishiyama, Y. (2011). Finite sample distributions of the J-B normality test statistics in dynamic models. Journal of Econometrics, 163(2), 149-157.

11. Henze, N., & Wagner, T. (1997). A new approach to the BHEP tests for multivariate normality. Journal of Multivariate Analysis, 60(1), 1-23.

12. Ferreira, M. A., & Steel, M. F. (2006). A constructive representation of univariate skewed distributions. Journal of the American Statistical Association, 101(474), 823-829.

13. Epps, T. W., & Pulley, L. B. (1983). A test for normality based on the characteristic function. Biometrika, 70(3), 723-726.

14. Switzer, P. (1985). A nonparametric test for gaussianity of stationary time series. Journal of Time Series Analysis, 6(3), 189-202.

15. Escanciano, J. C., & Velasco, C. (2006). A consistent diagnostic test for regression models using projections. Journal of Econometrics, 134(1), 1-25.

16. Greene, W. H. (1993). The econometric approach to efficiency analysis. In Empirical Economics, 18(1), 395-426.

17. Wang, J. J. (2007). Quantitative Assessment of Information Visualization Techniques. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2(2), 1-8.

18. Koenker, R. (2005). Quantile regression. In Econometric Society Monographs, 38, 1-3.

19. Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. In The American Statistician, 54(3), 217-224.

20. Shevlyakov, G. L., & Smirnov, P. O. (2011). A robust Jarque–Bera test. In Econometrics Journal, 14(1), 101-122.