ISSN PRINT 2319-1775 Online 2320-7876

 UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

Test Parameter Constancy and Predictive Performance in Statistical Models

Dr. A. Vidhyullatha[1], Dr. K.Dhana Lakshmi[2], Dr.B.Mamatha[3], Dr.K.Murali[4]**
[1]Assistant Professor in Mathematics, SPW Degree &PG College, TTD, Tirupati.
[2]Lecturer in Mathematics, SPW Degree and PG College, TTD, Tirupati.
[3]Lecturer in Mathematics, SPW Degree and PG College, TTD, Tirupati.
[4]Academic Consultant, Dept of Statistics ,S.V. University, Tirupati,
dr.murali4stat@gmail.com

ABSTRACT

Statistical models are fitted for various purposes. One key purpose is to observe relationships between variables. Another significant reason is to facilitate predictions, often used in selection processes. After fitting a regression model from a sample of observations, one may focus on predicting the value of the dependent variable for a specific value of an independent variable. This specific value of the independent variable might lie within the range of sample values, or more frequently, it might lie outside the sample observations.

A crucial criterion for an estimated regression equation is its relevance to data outside the sample used for estimation. This criterion is captured by the concept of parameter constancy, which means that the parameter vector should apply both within and outside the sample data. Parameter constancy can be assessed by testing predictive accuracy.

This research paper proposes tests for parameter constancy and predictive accuracy using different parameter vectors in the forecast period

I. INTRODUCTION

Ordinary Least Squares (OLS) regression is a cornerstone technique in statistical analysis, widely employed to investigate linear relationships between variables. Despite its extensive application, the method hinges on several critical assumptions, including the expectation that error terms have

ISSN PRINT 2319-1775 Online 2320-7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

a zero mean and constant variance. However, these assumptions are frequently violated in practical scenarios, leading to issues such as heteroscedasticity, where the error terms exhibit non-constant variance. These violations can undermine the efficiency and reliability of OLS estimates, making it crucial to employ diagnostic tools to identify and address potential problems.

Among the various diagnostic techniques available, the analysis of residuals stands out as a powerful method for detecting deviations from OLS assumptions. Two significant types of residuals used in this context are Studentized residuals and predicted residuals. Studentized residuals are particularly useful for identifying outliers and influential observations, as they standardize residuals by taking into account the variance of each individual residual. Predicted residuals, on the other hand, represent the differences between observed and predicted values of the dependent variable and can be instrumental in assessing the model's fit and predictive accuracy.

In situations where the OLS assumptions are violated, the standard OLS estimates may still be unbiased but lose efficiency, with biased standard errors. This necessitates the use of alternative methods, such as employing Studentized or predicted residuals, to enhance the robustness of the analysis. Additionally, when dealing with discrete dependent variables, OLS may not be the most suitable method, prompting the use of alternative models like probit or logit.

Understanding and implementing these diagnostic tools is crucial for researchers aiming to ensure the validity and reliability of their regression models. By carefully analyzing residuals, researchers can better detect and correct issues such as heteroscedasticity, outliers, and influential observations, ultimately leading to more accurate and dependable statistical analyses. This review of literature explores the significance and application of Studentized and predicted residuals in enhancing the diagnostic process for OLS regression models.

ISSN PRINT 2319-1775 Online 2320-7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

II. REVIEW OF LITERATURE

Ordinary Least Squares (OLS) regression is a foundational statistical method for examining linear relationships between variables (Kuchibhotla et al., 2018). Despite its widespread use, the method relies on several key assumptions, such as the error terms having a zero mean and constant variance. However, these assumptions are frequently violated in practice (Chapter 4 Regression Models, 2002). A common violation is heteroscedasticity, where the error terms' variance is not constant (Shelton, 1987).

To address these issues, various diagnostic tools have been developed, with the analysis of residuals being a prominent method. Studentized residuals and predicted residuals are particularly useful for identifying potential violations of OLS assumptions (Spector & Mazzeo, 1980).

Studentized residuals, also known as externally studentized residuals, are standardized to account for the variance of each individual residual, which can help in detecting outliers or observations with high influence on the regression results (Chapter 4 Regression Models, 2002). On the other hand, predicted residuals represent the differences between observed and predicted values of the dependent variable (Shelton, 1987).

When OLS assumptions are violated, the estimates may remain unbiased but lose efficiency, and the standard errors may become biased (Kuchibhotla et al., 2018). To mitigate these issues, researchers have proposed using studentized and predicted residuals. Studentized residuals are calculated by dividing the raw residuals by their estimated standard errors, aiding in the identification of outliers or influential observations. Predicted residuals, which are the differences between observed and predicted values, can be used to evaluate the model's goodness of fit and identify predictive inaccuracies.

Despite its popularity, OLS is not always suitable, particularly when dealing with discrete dependent variables. In such cases, alternative models like probit or logit are more appropriate (Spector & Mazzeo, 1980).

ISSN PRINT 2319-1775 Online 2320-7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

To improve the reliability of regression models, researchers employ diagnostic tools and techniques, such as the analysis of residuals. Studentized and predicted residuals are critical for assessing model fit and detecting outliers or influential observations. Internally studentized residuals are obtained by dividing the raw residual by its estimated standard error, following a t-distribution under normality and homoscedasticity assumptions, thus facilitating outlier detection. Externally studentized residuals are calculated by excluding the influence of each observation on the regression coefficients before computing the residual, helping identify influential observations (Assaf & Tsionas, 2019).

In summary, analyzing OLS, studentized, and predicted residuals is essential for validating regression models and identifying issues such as heteroscedasticity, outliers, and influential observations (Chatterjee & Wiseman, 1983; Dasgupta & Mishra, 2004). Careful consideration of these diagnostic measures is crucial for ensuring the reliability and validity of regression analysis findings. The residual analysis plays a vital role in assessing the adherence to OLS assumptions and detecting potential issues.

III. SOME IMPORTANT TYPES OF RESIDUALS

In statistical analysis, important types of residuals include standardized residuals, which are scaled by their standard deviation to detect outliers; studentized residuals, adjusted for leverage to assess influence; and deleted residuals, obtained by excluding each observation to evaluate its impact. Pearson residuals, primarily used in logistic regression, measure the difference between observed and predicted log odds, while deviance residuals in generalized linear models compare the model's likelihood to that of a saturated model. Cook's distance combines leverage and residual size to gauge the influence of observations on regression coefficients. Internal and external residuals address discrepancies within a model's data and in relation to external data, respectively.

ISSN PRINT 2319-1775 Online 2320-7876

 UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

IV. The OLS, Studentized and Predicted Residuals

Consider the Classical Linear Regression model as

$$y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \varepsilon_{n\times 1}$$

Such that $E[\varepsilon] = 0$ and $E[\varepsilon\varepsilon^I] = \sigma^2 I_n$. Here, the error vector ε consists of unknown errors which are assumed to be uncorrelated. Generally, ε can be estimated by the Ordinary Least Squares Residual vector e, which is given by

$$e = y - X\hat{\beta}$$

Where $\hat{\beta} = (X^I X)^{-1} X^I y$ is the Best Linear Unbiased Estimator (BLUE) of β

One may have,

$$\hat{y} = X\hat{\beta} = X[(X^I X)^{-1} X^I y]$$

and $$e = [y - \hat{y}] = [y - X(X^I X)^{-1} X^I y]$$

$$e = [I - X(X^I X)^{-1} X^I]y$$

$$e = [I - V]y$$

where $$V = X(X^I X)^{-1} X^I$$

Generally, $V = X(X^I X)^{-1} X^I$ is an important matrix and is known as HAT matrix. One may write,

$$y = [V + I - V]y = Vy + (I - V)y$$

Mean and Variance of e are given by

$$E[e - \varepsilon] = E[(I - V)y - \varepsilon]$$

ISSN PRINT 2319-1775 Online 2320-7876

 UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

$$= E[(I - V)(X\beta + \varepsilon) - \varepsilon]$$

$$= E[(I - V)\varepsilon - \varepsilon], [\because (I - V)X = 0]$$

$$= E[M - I]\varepsilon,$$

Where $M = [I - X(X^I X)^{-1} X^I]$ is a Symmetric Idempotent matrix,

$$\Rightarrow E[e - \varepsilon] = (M - I)E[\varepsilon] = 0$$

or $E[e] = 0 \quad [\because E(\varepsilon) = 0]$

Thus, the Ordinary Least Squares Residual vector e is a Linear Unbiased Estimator of ε. Also,

$$var(e) = E[ee^I]$$

$$= E[M\varepsilon\varepsilon^I M^I]$$

$$= ME[\varepsilon\varepsilon^1]M^I$$

$$= \sigma^2 MM^I$$

or $var(e) = \sigma^2 M$

Here, x_i^I and x_j^I are the i[th] row and j[th] row of the data matrix X respectively.

Now, the variance, covariance and correlation coefficients of the Residuals are given by,

$$var(e_i) = \sigma^2(1 - v_{ii})$$

$$cov(e_i e_j) = -\sigma^2 v_{ij}$$

$$r_{e_i e_j} = \frac{-v_{ij}}{\sqrt{(1 - v_{ii})(1 - v_{jj})}}$$

ISSN PRINT 2319-1775 Online 2320-7876

 UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

Also, V is (nxn) symmetric Idempotent matrix and Rank (V) = Rank (X) = k, k<n, where k is the number of parameters including intercept parameter and

$$Trace(V) = \sum_{i=1}^{n} v_{ii} = Rank(X) = k$$

Further, One may have, $\sum_j v_{ij}^2 = v_{ii}$

and $\sum_i v_{ij} = \sum_j v_{ij} =$

Each v_{ii} must fall in the interval, $0 \leq v_{ii} \leq 1$

The notion of v_{ii} measures the distance from the point x_i to the center of the data, and cases with unusual values for the independent variables will tend to have large values of v_{ii}

Generally, $var(e_i)$will be small whenever v_{ii} is large, so cases with X_i near $\bar{X}$ will be fit poorly and cases with x_i far from $\bar{X}$will be fit well, which is undesirable

Studentized residuals, an improved set of residuals, are obtained by scaling, so cases with larger v_{ii} get larger scaled residuals and those with smaller v_{ii} get smaller scaled residuals. This scaling is achieved by dividing each residual by an estimate of its standard deviation, making these residuals standardized versions that do not depend on σ^2sigma$\sigma^2$and the v_{ii} scale quantities. Margolin (1977) introduced the term "Studentized residuals" instead of "Standardized residuals." David (1981) identified two types of Studentized residuals: (i) Internally Studentized Residuals and (ii) Externally Studentized Residuals.

The Internally Studentized Residuals are given by,

$$e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1 - v_{ii}}}, i = 1,2 \cdots, n$$

ISSN PRINT 2319-1775 Online 2320-7876

 UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

Under the assumption of the normality of error vector ε in the Linear Regression model $y = X\beta + \varepsilon$, say $\varepsilon \sim N(0, \sigma^2 I_n)$, the Internally Studentized Residuals $\left[\frac{e_i^{*2}}{n-k}\right]$ follows Beta distributions with parameters $\left[\frac{1}{2}, \frac{n-k-1}{2}\right]$. One may obtain, $E[e_i^*] = 0$ and $var[e_i^*] = 1$ and

$$cov[e_i^*, e_j^*] = \frac{-v_{ij}}{\sqrt{(1-v_{ii})(1-v_{jj})}}, i \neq j = 1,2, \cdots, n$$

The Externally Studentized Residuals are given by

$$e_i^{**} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-v_{ii}}}, i = 1,2 \cdots, n$$

where, $\hat{\sigma}^2_{(i)} = \frac{\left[(n-k^I)\hat{\sigma}^2 - \frac{e_i^2}{1-v_{ii}}\right]}{n-k^I-1}$

or $\hat{\sigma}^2_{(i)} = \hat{\sigma}^2\left[\frac{n-k^I-e_i^{*2}}{n-k^I-1}\right]$

and $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}$

Here, Rank (V) = k^1

$\hat{\sigma}^2_{(i)}$ is an estimate of error variance σ^2 with i^{th} observation has been omitted from the estimation. Also, e_i^{**} follows t-distribution with (n-k^1-1) degrees of freedom. It can be shown that e_i^{**} is a monotonic transformation of e_i^* and a relationship between External and Internal Studentized residuals is given by,

$$e_i^{**} = e_i^*\left[\frac{n-k^I-1}{n-k^I-e_i^{*2}}\right]^{\frac{1}{2}}$$

In various diagnostic methods in Statistical Modelling, the predicted Residuals $e_{(i)}$ which is based on a fit to the data with the i^{th} observation excluded, has been frequently used. Suppose that $\hat{\beta}_{(i)}$ be the Ordinary Least Square

ISSN PRINT 2319-1775 Online 2320-7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

(OLS) estimator of β with i[th] observation has been excluded. Then the i[th] predicted Residual is defined by,

$$e_{(i)} = y_i - X_i^I \beta_{(i)}, i = 1,2,\cdots,n$$

Sometimes, $e_{(i)}$ may be known as prediction error.

The relationships among Ordinary least squares, Studentized and predicted Residuals are given by,

(i) $e_{(i)} = \frac{e_i}{1-v_{ii}}, i = 1,2,\cdots,n$

(ii) (ii) $e_i^* = \frac{e_{(i)}}{\frac{\hat{\sigma}}{\sqrt{(1-v_{ii})}}}$

(iii) $e_i^{**} = \frac{e_{(i)}}{\frac{\hat{\sigma}_{(i)}}{\sqrt{(1-v_{ii})}}}$

A widely used Criterion for Model selection known as "Predicted Residual Sum of Squares (PRESS)" based on $e_{(i)}$ is given by

$$PRESS = \sum e_{(i)}^2$$

Alternatively, interms of Studentized Residuals, one way express PRESS as

$$PRESS = \sum e_i^{*2}$$

or $PRESS = \sum e_i^{**^2}$

ISSN PRINT 2319-1775 Online 2320-7876

 UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

BIBLIOGRAPHY

1. Assaf, A G., & Tsionas, E G. (2019, June 26). Quantitative research in tourism and hospitality: an agenda for best-practice recommendations. Emerald Publishing Limited, 31(7), 2776-2787. https://doi.org/10.1108/ijchm-02-2019-0148
2. Brown, R.L., Durbin, J., and Evans, J.M. (1975), "Techniques for testing the constancy of Regression Relationships over time", Journal of the Royal Statistical Society Series-B, 37, 149-192.
3. Chambers, M. J., & McGarry, J. S. (2002), "Modeling cyclical behavior with differential-difference equations in an unobserved components framework", Econometric Theory, 18(2), 387-419.
4. Chätterjee, S., & Wiseman, F. (1983, August 1). Use of Regression Diagnostics in Political Science Research. Wiley, 27(3), 601-601. https://doi.org/10.2307/2110986
5. Chu, C J., Stinchcombe, M B., & White, H. (1996, September 1). Monitoring Structural Change. Wiley, 64(5), 1045-1045. https://doi.org/10.2307/2171955
6. Cox, D.R. (1961), "Tests of Separate Families of Hypotheses", in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.
7. Dasgupta, M., & Mishra, S K. (2004, January 1). Least Absolute Deviation Estimation of Linear Econometric Models: A Literature Review. RELX Group (Netherlands). https://doi.org/10.2139/ssrn.552502
8. Honda, T. (1997, January 1). THE CUSUM TESTS WITH NONPARAMETRIC REGRESSION RESIDUALS. Japan Statistical Society, 27(1), 45-63. https://doi.org/10.14490/jjss1995.27.45
9. Kadane, J.B., and Lazar, N.A. (2004), "Methods and Criteria for Model Selection", Journal of the American Statistical Association, 99, 465, 279-290.

ISSN PRINT 2319-1775 Online 2320-7876

 UGC CARE Listed (Group -I) Journal Volume 11, Iss 3, March 2022

10. Kuchibhotla, A K., Brown, L D., & Buja, A. (2018, January 1). Model-free Study of Ordinary Least Squares Linear Regression. Cornell University. https://doi.org/10.48550/arxiv.1809.10538

11. Lang, S., Adebayo, S., and Fahremir, L. (2002), "Bayesian Semiparametric Seemingly Unrelated Regression", Proceedings in Computational Statistics (ed. By W. Hardle and B. Ronz), Physika- Verlag, Heidel berg, (195-200).

12. Shelton, F A. (1987, January 1). Using regression analysis: A guided tour. Elsevier BV, 11(2), 95-111. https://doi.org/10.1016/0360-1315(87)90004-2

13. Spector, L., & Mazzeo, M J. (1980, March 1). Probit Analysis and Economic Education. Taylor & Francis, 11(2), 37-44. https://doi.org/10.1080/00220485.1980.10844952

14. Turner, P. (2010, July 16). Power properties of the CUSUM and CUSUMSQ tests for parameter instability. Taylor & Francis, 17(11), 1049-1053. https://doi.org/10.1080/00036840902817474

15. Wright, J. (1993, January 1). The CUSUM test based on least squares residuals in regressions with integrated variables. Elsevier BV, 41(4), 353-358. https://doi.org/10.1016/0165-1765(93)90204-p