International Journal of Mathematical Archive-5(3), 2014, 87-94
Available online through www.ijma.info ISSN 2229 – 5046

A MODIFIED SHAPIRO – FRANCIA TEST FOR NORMALITY OF DISTURBANCES IN LINEAR STATISTICAL MODEL

Vijaya Kumar K*[1] and Balasiddamuni P[2]

[1]Department of Statistics, S. G. S. Arts College, TTD, Tirupati, Andhra Pradesh, India.

[2]Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh, India.

(Received on: 07-03-14; Revised & Accepted on: 24-03-14)

ABSTRACT

In the present study, a procedure has been discussed for testing normality of disturbances in linear statistical model. A modified version of Shapiro-Francia Test for normality of disturbances based on internally studentized residuals has been proposed as large sample test.

1. INTRODUCTION

In many empirical studies using the linear regression model, it is standard practice to assume that the random disturbances have zero means, and are normally, independently and identically distributed. Since, failure of these assumptions to hold can lead to biased and inconsistent estimators; and incorrect inferences, it is sensible to test these assumptions rather than to presume that they are correct.

Tests of normality are statistical inference procedures designed to test that the underlying distribution of a random disturbance variable is normally distributed. There is a long history of these tests, and there are a plethora of them available for use. The recent interest in goodness-of-fit tests for normality has arisen from the exciting work of Shapiro and Wilk (1965).

The problem of non-normal disturbances is one of the serious problems in the theory of Econometrics. The presence of non-normal disturbances disturbs the optimal properties of the Ordinary Least Squares (OLS) estimators of the parameters of a linear statistical model. Hence, there is a necessity of detecting the problem of non-normal disturbances so that the various techniques can be applied to avoid the presence of non-normal disturbances in the linear statistical model.

In the present study, an attempt is made to propose a modified version of Shapiro-Francia test statistic based on Internally studentized residuals, for testing normality of disturbances in linear statistical models.

2. REVIEW OF THE LITERATURE

The main contribution relating to the problem of testing normality of the disturbances in the linear statistical models, has been made by Geary (1935, 1947), Shapiro and Wilk (1965), Shapiro, Wilk and Chen (1968), Shapiro and Francia (1972), Huang and Bolch (1974), Filliben (1975), Weisberg and Bingham (1975), Locke and Spurrier (1976), Gastwirth and Owens (1977), Spiegelhather (1977), Pearson, D'A gostino and Bowman (1977), White and Mac Donald (1980), Pierce and Gray (1982), Royston (1982), La Riccia (1986), Jarque and Bera (1987), Bonett and Woodward (1990), Kalirajan and Jayasuriya (1991), Al-shiha and Yang (1996), Markatou and Manos (1996) and Rahman and Govindarajulu (1997).

Most of the tests for normality discussed in the literature have used either the Ordinary Least Squares (OLS) residual vector (e) or the best linear unbiased scalar (BLUS) residuals vector $(\tilde{e})$ or the recursive residual vector $(\tilde{e})$.

Corresponding author: Vijaya Kumar K[1]*
[1]Department of Statistics, S. G. S. Arts College, TTD, Tirupati, Andhra Pradesh, India.
E-mail: vijayneeraja73@gmail.com

SHAPIRO AND FRANCIA W[1] TEST FOR NORMALITY

Shapiro and Wilk (1965) suggested w test statistic for normality as

$$w = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{2.1}$$

Where, $a^{|} = \left[a_1 a_2 .....a_n\right] = \frac{m^{|} v^{-1}}{\left[m^{|} v^{-|} v^{-|} m\right]^{1/2}}$ (2.2)

$V = \left(\left(V_{ij}\right)\right)$ is (n x n) covariance matrix;

$m^{|} = \left[m_1 m_2 .....m_n\right] =$ vector of expected values of standard normal order statistics.

$y^{I} = \left[y_1 y_2 .....y_n\right] =$ vector of ordered random observations

The elements of V are given only for sample sizes upto 20 and approximations were developed by Shapiro and Wilk (1965) for calculating coefficients $\{a_i\}$ for using the w test statistic upto the sample size 50.

The overcome the main drawback of Shapiro and Wilk (1965) W test statistic for large samples, the observation $\{y_i\}$ may be treated as if they were independent and the identify matrix I_n can be substituted for V^{-1} in the estimation of the slope of the regression line

$$y_i = \mu + \sigma X_i,\ i = 1,2,3.....n. \tag{2.3}$$

Consequently, Shapiro and Francia (1972) suggested an approximate $W^{|}$ test statistic for normality as

$$w^{|} = \frac{\left(\sum_{i=1}^{n} b_i y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{2.4}$$

Where, $b^{|} = \left[b_1\ b_2\ .....b_n\right] = \frac{m^{|}}{\sqrt{m^{|} m}}$ (2.5)

Values of 'm' are given for different sample sizes by Harter (1961). The null distribution of $W^{|}$ statistic was approximated by an empirical sampling study. Shapiro and Francia (1972) conducted an empirical sampling study to compare the sensitivities of the W and $W^{|}$ test statistics in detecting non-normality. The empirical percentage points of $W^{|}$ test statistic for different sample sizes were tabulated and concluded that in general, the $W^{|}$ test appears to be more sensitive than the W test when the alternative distribution was continuous and symmetric with high fourth moment (as compared to the normal distribution); when it was near normal and when it was discrete and skewed. The two tests W and $W^{|}$ appeared to be equivalent for alternative distributions which were continuous and asymmetric with high fourth moment and discrete and symmetric. The W test is superior to the $W^{|}$ test for other alternative distributions. Overall, the differentials in the power were small.

3. STUDENTIZATION OF RESIDUALS

By Studentization, the OLS residuals can be transformed to have a selected covariance structure and the null distributions of the transformed residuals are independent of the scale parameters.

Since, $Var(e) = \sigma^2(I - V)$ *or* $Var(e_i) = \sigma^2(I - v_{ii}) \forall\ i = 1,2,3,....n.$, the OLS residuals have the distribution which is scale dependent. Margolin (1977) defined the studentized residuals as the division of a scale dependent

statistic say U, by a scale estimate T, so that the ratio $S = \frac{U}{T}$ has a distribution that is free of the nuisance scale parameters.

David (1981) distinguished two type of studentization of residuals namely

(i) Internal studentization, where U and T are dependent and
(ii) External studentization, where U and T are independent.

(A) INTERNAL STUDENTIZATION OF RESIDUALS

Under internal studentization, the 'Internally studentized Residuals' are

$$e_i^+ = \frac{e_i}{\hat{\sigma}\sqrt{1-v_{ii}}}, i = 1, 2, ..., n. \tag{3.1}$$

Where $\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-k}}$ (3.2)

Beckman and Trussell (1974) have proved that, $e_i^{+^2} \Big/ n-k$ follows a Beta distribution with prameters $\frac{1}{2}$ and $\frac{n-k-1}{2}$. The probability density function of e_i^+ is given by

$$f(e_i^+) = \frac{\Gamma\left(q+\frac{1}{2}\right)}{\Gamma(q)\Gamma\left(\frac{1}{2}\right)\sqrt{n-k}}\left(1-\frac{e_i^{+^2}}{n-k}\right)^{q-1}; \tag{3.3}$$

and $\left|e_i^+\right| \le \sqrt{n-k}$

Here, $q = \frac{n-k-1}{2}$

It follows that $E\left[e_i^+\right] = 0, \ \forall\ i = 1, 2, 3, ...n$

$Var\left(e_i^+\right) = 1;$ and

$$\operatorname{cov}\left(e_i^+ e_j^+\right) = \frac{-v_{ij}}{\sqrt{\left(1-v_{ii}\right)\left(1-v_{jj}\right)}}, \ \forall\ i \neq j \tag{3.4}$$

(B) EXTERNAL STUDENTIZATION

By external studentization, the 'Externally Studentized Residuals' are defined by

$$e_i^{++} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-v_{ii}}} \tag{3.5}$$

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

or $\hat{\sigma}_{(i)}^2 = \hat{\sigma}^2\left[\frac{n-k-e_i^{+^2}}{n-k-1}\right]$ (3.6)

Here, $\hat{\sigma}_{(i)}^2$ is the residual mean square computer without i[th] observation on both Y and X's. Under normality, $\hat{\sigma}_{(i)}^2$ and e_i are independent.

The probability density function of e_i^{++} is the Student's t-distribution with (n-k-1) degree of freedom.

A relationship between e_i^{++} and e_i^{+} is given by

$$e_i^{++} = e_i^{+}\left[\frac{n-k-1}{n-k-e_i^{+^2}}\right]^{1/2} \tag{3.7}$$

This shows that $e_i^{++^2}$ is a monotonic transformation of $e_i^{+^2}$.

Studentized residuals provide way to examine the information in the residuals, both because they have equal variances and because they are easily related to the t-distribution in many situations.

4. A MODIFIED SHAPIRO – FRANCIA TEST FOR NORMALITY OF DISTURBANCES

Consider the classical linear regression model,

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \in_{n\times 1} \tag{4.1}$$

Where, X is a known nonstochastic (nxk) matrix of rank k; β is a (kx1) vector of unknown parameters: Y is the (nx1) vector of random observation; and $\in$ is an (nx1) random vector whose elements are assumed to be normally distributed with mean zero and constant variance σ^2.

Since, the disturbance vector $\in$ is unobservable, the various tests such as tests for specification errors in the model, tests for normality etc., generally rely on sample residuals such as OLS or BLUS or Recursive or studentized or predicted residuals. However, OLS residuals are considered to be inferior to other residuals by many econometricians, since the OLS residuals vector has a singular normal distribution and the elements of the vector are not independently and identically distributed.

The OLS residual vector e associated with the true disturbance vector $\in$ is given by

$$e = M\in \tag{4.2}$$

where, $M = \left[I - X(X^{|}X)^{-|}X^{|}\right]$ is an (nxn) symmetric idempotent matrix with rank (n-k). Thus e has a singular normal distribution with zero mean vector and covariance matrix is given by $\sigma^2 M$.

The application of the tests to the disturbances offers a set of benchmark results for gauging the performance of the test as applied to identically and independently distributed random variables.

The recent interest in goodness of fit tests for normality has arisen from the exciting work of Shapiro and Wilk (1965),

The (n-k) x 1 matrix of BLUS residuals is given by

$$\tilde{e} = A^{|}\in \tag{4.3}$$

Where $A^{|}$ is (n-k) xn such that $A^{|}A = I_{n-k}$ and

$E\left[\tilde{e}\tilde{e}^{|}\right] = \sigma^2 I_{n-k}$. *Let* $A = \left(\left(a_{ij}\right)\right)$

Thus, the BLUS residuals are linear combinations of the disturbance and a set of coefficients such that

$$\left.\begin{aligned}\sum_{r=1}^{n} a_{ri}a_{rj} &= 0\\ \sum_{r=1}^{n}\left(a_{ri}a_{rj}\right)^2 &\neq 0\end{aligned}\right\} \forall\ i \neq j \tag{4.4}$$

By applying the following theorem given by Kingman and Grabill (1970):

Let $Y_1, Y_2, \ldots\ldots, Y_n$ be i.i.d. random variables with cdf denoted by F and let n be a positive integer. Suppose that L_1 and L_2 are defined by

$$L_1 = \sum_{i=1}^{n} a_i Y_i \; ; \qquad L_2 = \sum_{i=1}^{n} b_i Y_i \tag{4.5}$$

Where a_i and b_i are constants such that

$$\sum_{i=1}^{n} a_i b_i = 0 \quad \text{and} \qquad \sum_{i=1}^{n} (a_i b_i)^2 \neq 0$$

Then F is the cdf of a normally distributed random variable if and only if L_1 and L_2 are independent".

Using this theorem, we find that if the disturbances are independently and identically distributed, the BLUS residuals can be independent if and only if the disturbances are normally distributed. BLUS residual vector $\tilde{e}$ has a nonsingular normal distribution with a covariance matrix given by $\sigma^2 I_{n-k}$.

Suppose $\tilde{e}_{(1)} \leq \tilde{e}_{(2)} \leq \ldots\ldots \leq \tilde{e}_{(n-k)}$ denote an ordered random sample of (n-k) BLUS residuals derived from n OLS residuals.

Let $m^{|} = (m_1, m_2 \ldots\ldots\ldots, m_{n-k})$ be the vector of expected values of ordered BLUS residuals and

$$b^{|} = \frac{m^{|}}{\sqrt{m^{|} m}}, \text{ where } b^{|} = (b_1, b_2 \ldots\ldots\ldots, b_{n-k}) \tag{4.6}$$

By using Shapiro and Francia (1972) approximate Analysis of variance test of normality, a modified test statistic for normality is given by

$$W^{*} = \frac{\left[\sum_{i=1}^{n-k} b_i \tilde{e}_{(i)}\right]^2}{\sum_{i=1}^{n-k} \left[\tilde{e}_{(i)} - \bar{\tilde{e}}\right]^2} \tag{4.7}$$

where, $\bar{\tilde{e}} = \dfrac{\sum_{i=1}^{n-k} \tilde{e}_i}{n-k}$ (4.8)

This test can be performed by using the critical values tabulated by Shapiro and Francia (1972)

Two more modification of test statistic, using the approximations suggested by Weisburg and Binham (1975) and Blom (1956) are given by:

(a) $$b^{*|} = \frac{m^{*|}}{\sqrt{m^{*|} m^{*}}} \tag{4.9}$$

where $$m^{*} = \phi^{-|}\left(\frac{i - 3/8}{n + 1/4}\right), \quad i = 1, 2, \ldots, (n-k) \tag{4.10}$$

and $\phi^{-|}$ denotes the inverse of the standard normal cumulative distribution function. Then the test statistic is given by

$$W^{**} = \frac{\left[\sum_{i=1}^{n-k} b_i^{*} \tilde{e}_{(i)}\right]^2}{\sum_{i=1}^{n-k} \left[\tilde{e}_{(i)} - \bar{\tilde{e}}\right]^2} \tag{4.11}$$

(b) $b^{**|} = \dfrac{m^{**|}}{\sqrt{m^{**|}\ m^{**}}}$, (4.12)

where, $m^{**} = \phi^{-|}\left(P_i\right), \quad i = 1, 2, \ldots, (n-k)$ (4.13)

Here, $P_i = \dfrac{i}{n+1}, \ i = 1, 2, 3, \ \ldots, (n-k)$

$$W^{***} = \frac{\left[\sum_{i=1}^{n-k} b_i^{**}\, \tilde{e}_{(i)}\right]^2}{\sum_{i=1}^{n-k}\left[\tilde{e}_{(i)} - \bar{\tilde{e}}\right]^2} \tag{4.14}$$

The main advantages of these test statistics that they remove the need to have tables of weights for computation of the test statistics.

The proposed modified test statistic will provide Goodness of fit test for normality for all sample sizes without the assumption of zero-correlation among the ordered residuals.

It should be noted that, although the above tests assume that the residuals are independent, both Haung and Bolch (1974) and Ramsey (1974) reported from their Monte Carlo studies, where the correlated OLS residual vector led to a more powerful test than that obtained using the BLUS residual vector.

Now, we propose a modified Shapiro-Francia large sample test based on Internally studentized residual vector e^+ as follows.

$$W^+ = \frac{\left[\sum_{i=1}^{n} b_i^+\ e_{(i)}^+\right]^2}{\sum_{i=1}^{n} e_{(i)}^{+^2}} \tag{4.15}$$

where, $b^{+|} = \dfrac{m^{+|}\,\Sigma^{-|}}{\sqrt{m^{+|}\,\Sigma^{-|}\,m^+}}$; (4.16)

$m^{+|} = \left(m_1^+, m_2^+, ..., m_n^+\right)$ = Vector of expected values of ordered Internally studentized residuals;

$e_{(1)}^+ \le e_{(2)}^+ \le ... \le e_{(n)}^+$ are the ordered studentized residuals; $\sum\left(\left(\sigma_{ij}\right)\right)$ is the nxn covariance matrix of ordered Internally studentized residuals; and

$$\sigma_{ij} = 1 \ \forall\, i = j = 1, 2, ..., n$$

$$= \frac{-\nu_{ij}}{\sqrt{\left(1-\nu_{ii}\right)\left(1-\nu_{jj}\right)}} \ \forall\ i \neq j = 1, 2, ..., n \tag{4.17}$$

Here, $V = \left(\left(\nu_{ij}\right)\right) = X\left(X^{|}X\right)^{-|}\ X^{|}$ = Hat matrix (4.18)

Under the following assumptions given by white and Mac Donald (1980) for the large sample behavior of the modified tests for nonnormality is the linear regression model, it can be shown that the prosposed modified Shapiro-Francia Large sample test statistic W^+ is consistent estimator of the true statistic W.

Assumption (i) The linear regression model is

$$Y_i = X_i^{|}\beta + \in_i, \quad i = 1, 2, \ldots, n \tag{4.19}$$

Where $|X_i|$ is a sequence of uniformly bounded fixed (kx1) vector such that

$$\frac{X^{|}X}{n} \to Q_{xx}, \text{ a positive definite matrix} \tag{4.20}$$

$|\in_i|$ is a sequence of i.i.d. random variables

$$\left.\begin{aligned} \text{with} \quad & E\left[\in_i\right] = 0, \ \forall i = 1, 2, \ldots, n \\ & E\left[\in_i^2\right] = \sigma^2, \ 0 < \sigma^2 < \infty \end{aligned}\right\} \tag{4.21}$$

β is a (kx1) vector of parameters

This is assumption is sufficient to ensure

$$\hat{\beta} \xrightarrow{a.s} \beta \tag{4.22}$$

Assumption (ii) (a) The density of $\in_i$, say f = dF is uniformly continuous, positive on the interval of support and bounded;

(b) Define

$$\left.\begin{aligned} X_{ip}^{+} &= \text{Max}\left\{X_{ip}, 0\right\} \quad \text{and} \\ X_{ip}^{-} &= X_{ip}^{+} - X_{ip}; \ i = 1, 2, \ldots, n \\ & \qquad p = 1, 2, \ldots, k \end{aligned}\right\} \tag{4.23}$$

$$\text{Max}\left\{X_{ip}^{+} : 1 \le i \le n\right\} = 0\left\{\left[\sum_{i=1}^{n}\left(X_{ip}^{+}\right)^2\right]^{1/2}\right\} \text{ and}$$

$$\text{Max}\left\{X_{ip}^{-} : 1 \le i \le n\right\} = 0\left\{\left[\sum_{i=1}^{n}\left(X_{ip}^{+}\right)^2\right]^{1/2}\right\}, \ p = 1, 2, \ldots, k$$

These two assumptions imply the necessary and sufficient conditions for the asymptotic normality of $\sqrt{n}\left(\hat{\beta} - \beta\right)$.

Under the above two conditions, it can be shown that

$$\left|W^{+} - W\right| \xrightarrow{p} 0 \tag{4.24}$$

Where W is Shapiro-Francia test statistic in terms of disturbances $\in_i$'s.

Small values of W^{+} are significant, that indicates nonnormality of disturbances. A more precise significance level may be associated with an observed W+ value by using the critical values given by Shapiro and Francia (1972).

5. CONCLUSIONS

In the present study a modified version of Shapiro-Francia test for normality of disturbances in a linear statistical model has been discussed by using Internally Studentized Residuals. This proposed test statistic is a consistent estimator of the true test statistic under the certain assumptions given by White and Mac Donald (1980). This kind of study can be further extended by deriving power functions of the tests, for the comparative study of the various tests of normality of disturbances.

REFERENCES

1. Al-shiha, A.A. and Yang, S.S. (1996), "An Improved Approximation for the Shapiro-Wilk Test Statistic for Normality", Pakisthan Journal of Statistics, 12,215-230.
2. Beckman, R. and Trussell, H. (1974), "The Distribution of an Arbitrary Studentized Residual and the Effects of updating in Multiple Regression", JASA, 69, 199-201.
3. Blom, G. (1956), "Statistical Estimates and Transformed Beta Variables", New York: John Wiley.
4. Bonett, D.G. and Woodward, J.A. (1990), "Testing Residual Normality in the ANOVA Model", Journal of Applied Statistics, 17, 383-387
5. David, H.A. (1981), Order Statistics, 2nd Edition Wiley, New York.
6. Filliben, J.J. (1975), "The Probability plot Correlation Co-efficient Test for Normality", Technometrics, 17, 111-117.
7. Gastwirth, J.L. and Owens, M.E.B.(1977), "On Classical Tests of Normality", Biometrika, 64, 135-139
8. Geary, R.C. (1935), "Note on the Correlation Between $p_z$and W^1", Biometrika, 27, 353-355.
9. Geary, R.C. (1947), "Testing for Normality", Biometrika, 34, 209-242.
10. Harter, H.L. (1961), "Expected Values of Normal Order Statistics", Biometrika, 48, 151-165.
11. Huang, C.J. and Bolch, B.W. (1974), "On the Testing of Regression Disturbances for Normality", JASA, 69, 330-335.
12. Jarque, C.M. and Bera, A.K. (1987), "A Test for Normality of Observations and Regression Residuals", International Statistical Review, 55, 163-172.
13. Kalirajan, K.P. and JayaSuriya, S.K.W. (1991), "Simultaneous Testing of Regression Disturbances for Heteroscedasticity and Non-Normality", Journal of Applied Statistics, 18, 307-312.
14. Kingman, A. and Gray Bill, E.A. (1970), "A Nonlinear Characterization of the Normal Distributon", The Annals of Mathematical Statistics, 41, 1889-95.
15. La Riccia, V.N. (1986), "Optimal Goodness-of-Fit Tests for Normality Against Skewness and Kurtosis Alternatives", Journal of Statistical Planning and Inference, 13, 67-69.
16. Locke, C. and Spurrier, J.D. (1976), "The Use of U-Statistics for Testing Normality Against Non-symmetric Alternatives", Biometrika, 63, 143-147.
17. Margolin, B.H. (1977), "The Distribution of Internally Studentized Statistics Via Laplace Transform Inversion", Biometrika, 64, 573-582.
18. Markatou, M. and Manos, G. (1996), "Robust Tests in Nonlinear Regression Models", Journal of Statistical Planning and Inference, 55, 205-217.
19. Pearson, E.S., D'Agostino, R.B. and Bowman, K.O. (1977), "Tests for Departure for Normality: Comparision of Powers", Biometrika, 64, 231-246.
20. Pierce, D.A. and Gray, R.J. (1982), "Testing Normality of Errors in Regression Models", Biometrika, 69, 233-236
21. Rahman, M.M. and Govindarajulu, Z. (1997), "A Modification of the Test of Shapiro and Wilk for Normality", Journal of Applied Statistics, 24, 219-235.
22. Ramsey, J.B. (1974), "Classical Model Selection Through Specification Error Tests", in P. Zarembka, ed., Frontiers in Econometrics, New York: Academic.
23. Royston, J.P. (1982), "An Extension of Shapiro and Wilk's W Test for Normality to Large Samples", Applied Statistics, 2, 115-124.
24. Shapiro, S.S. and Francia, R.S. (1972), "An Approximate Analysis of Variance Test for Normality", JASA, 67, 215-216.
25. Shapiro, S.S. and Wilk, M.B. (1965), "An Analysis of Variance Test for Normality (Complete samples)", Biometrika, 52, 591-611.
26. Shapiro, S.S., Wilk, M.B. & Chen, H.J. (1968), "A Comparative Study of Various Tests for Normality", JASA, 63, 1343-1372.
27. Spiegelhalter, D. J. (1977), "A Test for Normality against symmetric Alternatives", Biometrika, 64, 415-418.
28. Vijaya Kumar, K. (2013), "Nonnormal Disturbances; Testing Normality in Linear Statistical Models", Ph.D., Thesis, S.V. University.
29. Weisberg, S. and Bingham, C. (1975), "An Approximate Analysis of Variance Test for Non-Normality suitable for Machine Calculation", Technimetrics, 17, 133-134
30. White, H. and Mac Donald, G.M. (1980), "Some Large-Sample Tests for Nonnormality in the Linear Regression Model", JASA, 75, 16-28.

Source of support: Nil, Conflict of interest: None Declared

