Proceedings of the 1st ISM International Statistical Conference 2012, 560 – 565
4 – 6 September 2012, PERSADA Johor Bahru, Johor, Malaysia

A Test for Normality in the Presence of Outliers

Pooi Ah Hin[1] & Soo Huei Ching[2]
Sunway University Business School
No 5, Jalan Universiti, Bandar Sunway
46150 Petaling Jaya, Selangor, Malaysia
e-mail: [1]ahhinp@sunway.edu.my, [2]hueichings@sunway.edu.my

Abstract The Jarque-Bera test is a test based on the coefficients of skewness (S) and kurtosis (K) for testing whether the given random sample is from a normal population. When the random sample of size n contains m outliers, we use the remaining $n - m$ observations to compute two statistics S^* and K^* which mimic the statistics S and K. The statistics S^* and K^* are next transformed to z_1 and z_2 which are uncorrelated and having standard normal distributions when the original population is normal. We show that the acceptance region given by a circle in the $z_1 - z_2$ plane is suitable for testing the normality assumption.

Keywords The Jarque-Bera test, outliers

1 Introduction

The validity of many statistical procedures depends on normality assumption of observations. Several methods have been introduced in the literature for assessing the assumption of normality. The more popular ones among them are the Jargue-Bera test [1], Shapiro-Wilk test [2] and Anderson Darling test [3]. The Jargue-Bera (JB) test for normality is based on the sample coefficients of skewness and kurtosis, the Shapiro-Wilk test relies on the correlation in the Q-Q plot, and the Anderson-Darling test makes use of the difference between empirical and theoretical distributions.

The JB test has been modified by a number of authors [4-6]. Gel and Gastwirth [4] replaced the denominators of skewness and kurtosis in the JB test statistics by the average absolute deviation from the median (MAAD) to obtain the Robust Jargue-Bera (RJB) test statistic. In regression analysis, the estimates of the coefficients of skewness and kurtosis based on the Ordinary Least Squares (OLS) residuals have been rescaled by Imon [5], and the resulting JB type test is called the Rescalled Moment (RM) test. By replacing the estimate of spread with MAAD, the RM test can be modified to yield yet another test called the Robust Rescaled Moment (RRM) test [6].

Presently we consider the problem of assessing normality given a random sample $y_1, y_2, ..., y_n$ of size n and containing m outliers. We assume that the outliers are not due to careless mistakes. Instead we assume that the outlies are due to deficiencies in the measuring system in measuring units of which the attributes to be measured have very large or small values. For example, a spring balance may stretch disproportionately when the object to be weighed is very heavy. We thus assume that the extremely large (or small) values which are classified as outliers in the given sample are originally very large (or small) values but distorted to some extremely large (or small) values which are clearly not concordant with the rest of the data. With this assumption, we define the k-th adjusted sample moment by

$$m_k^* = \frac{1}{n-m} \sum_{i=m_1+1}^{n-m_2} (y_i - M)^k, \quad k = 1, 2, \ldots$$

where

m_1 is the number of outliers at the lower end,

m_2 the number of outliers at the upper end,

$m = m_1 + m_2$,

and M the sample median based on the original sample of size n. Hence the adjusted coefficients of skewness and kurtosis can be defined respectively as

$$S* = \frac{m_3^*}{[m_2^*]^{3/2}} \text{ and } K* = \frac{m_4^*}{[m_2^*]^2}$$

The statistics S^* and K^* are next expressed respectively as nonlinear functions of the uncorrelated random variables z_1 and z_2 which have standard normal distributions when the original population is normal. We show that the acceptance region given be a circle with centre zero in the z_1–z_2 plane is suitable for testing the normality assumption.

The layout of the paper is as follows. In Section 2, we state the Jargue-Bera test. In Section 3, we introduce a type of non-normal distribution called the quadratic-normal distribution [7]. In Section 4, we describe the method in [8] which makes use of the quadratic-normal distribution for transforming a set of non-normal random variables with a nonlinear dependence structure to a set of uncorrelated random variables which have standard normal distributions. The method in Section 4 will be used in Section 5 for transforming the statistics S^* and K^* which mimic the coefficients of skewness (S) and kurtosis (K) to the uncorrelated standard normal variables z_1 and z_2 . Section 6 compares the power of the test based on a circle in the $z_1 - z_2$ plane derived from (S *, K^*) when there are outliers with the power of the test based on a circle in the $z_1 - z_2$ plane derived from (S, K) when there are no outliers, and also with the power of the JB test. In Section 7, we give some concluding remarks.

2 Jarque-Bera Test

Given the random sample $\mathbf{y} = (y_1, y_2, ..., y_n)^T$, the sample coefficients of skewness and kurtosis are given respectively by

$S = \frac{m_3}{m_2^{3/2}}$ and $K = \frac{m_4}{m_2^2}$

where $m_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^k$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The Jargue-Bera test statistic is given by

$$\text{JB} = \frac{S^2}{\sigma_S^2} + \frac{(K-3)^2}{\sigma_K^2}$$

where $\sigma_S^2 = 6/n$ and $\sigma_K^2 = 24/n$ are respectively the asymptotic variances of S and K. When the y_i are normally distribution, the JB statistic has approximately a chi square distribution with 2 degrees of freedom. The rejection region, based on the above chi square approximation, for the hypothesis of normality is given by

$$\{\text{y: JB} > \chi^2_{2,\alpha}\}$$

where $\chi^2_{2,\alpha}$ is the $100(1-\alpha)\%$ point of the chi square distribution with 2 degrees of freedom.

3 Quadratic-normal Distribution

Let μ, $\lambda = (\lambda_1, \lambda_2, \lambda_3)^T$ be constants and consider the following nonlinear transformation

$$y = \begin{cases} \mu + \lambda_1 z + \lambda_2\left(z^2 - \frac{1+\lambda_3}{2}\right), & z \geq 0, \\ \mu + \lambda_1 z + \lambda_2\left(\lambda_3 z^2 - \frac{1+\lambda_3}{2}\right), & z < 0. \end{cases}$$

where z has the standard normal distribution and y is a one-to-one function of z. The random variable y is then said to have the quadratic-normal distribution with parameters μ and λ [7]. We may write $y \sim \text{QN}(\mu, \lambda)$.

4 Transforming correlated non-normal variables to uncorrelated standard normal variables

Let $\mathbf{t} = (t_1, t_2, ..., t_k)^T$ be a set of correlated non-normal random variables. Suppose we have a set of N_s observed values of $\mathbf{t}$. Let the n_s-th observed value be denoted by $(t_{1n_s}, t_{2n_s}, ..., t_{kn_s})^T$.

The following is a procedure modified from the method given in Pooi (2006) for fitting a non-normal distribution to the observed values of $\mathbf{t}$:

(i) Find the average value of t_i:

$$\bar{t}_i = \frac{1}{N_s}\sum_{n_s=1}^{N_s} t_{in_s}.$$

(ii) Estimate the moment $E\left[t_i^*\right]^{k_1} E\left[t_j^*\right]^{k_2}$ using $\hat{M}_{ij}^{(t)(k_1,\,k_2)} = \frac{1}{N_s}\sum_{n_s=1}^{N_s}\left(t_{in_s}^{*k_1} t_{jn_s}^{*k_2}\right)$, $k_1 \geq 0,\ k_2 \geq 0,\ k_1 + k_2 = 2$, where $t_{ij}^* = t_{ij} - \bar{t}_i$.

(iii) Compute the variance-covariance matrix, $\mathbf{M} = \left\{\hat{M}_{ij}^{(t)(1,\,1)}\right\}$ and find the matrix V formed by the eigenvectors of $\mathbf{M}$.

(iv) Find $\begin{bmatrix} \tilde{u}_{1n_s} \\ \tilde{u}_{2n_s} \\ \vdots \\ \tilde{u}_{kn_s} \end{bmatrix} = \mathbf{V}^T \begin{bmatrix} t_{1n_s}^* \\ t_{2n_s}^* \\ \vdots \\ t_{kn_s}^* \end{bmatrix}$.

(v) Compute the moment $\hat{M}_{ij}^{(\tilde{u})(k_1,\,k_2)} = \frac{1}{N_s}\sum_{n_s=1}^{N_s}\left(\tilde{u}_{in_s}^{k_1},\ \tilde{u}_{jn_s}^{k_2}\right)$, $k_1 \geq 0,\ k_2 \geq 0,\ 1 \leq k_1 + k_2 \leq 4$.

(vi) Find $u_{in_s} = \tilde{u}_{in_s} \Big/ \sqrt{\operatorname{Var}(\tilde{u}_i)}$ where $\operatorname{Var}(\tilde{u}_i) = \hat{M}_{ii}^{(\tilde{u})(1,\,1)}$.

(vii) Compute the moment $\hat{M}_{ij}^{(u)(k_1,\,k_2)} = \frac{1}{N_s}\sum_{n_s=1}^{N_s} u_{in_s}^{k_1} u_{jn_s}^{k_2}$, $k_1 \geq 0,\ k_2 \geq 0,\ 1 \leq k_1 + k_2 \leq 4$.

(viii) From the moment $\hat{M}_{ii}^{(u)(k,\,0)}$, find $\lambda^{(i)}$such that the k-th moment of a random variable ε_i with the QN$(0,\ \lambda^{(i)})$ distribution is equal to $\hat{M}_{ii}^{(u)(k,\,0)}$, $2 \leq k_1 \leq 4,\ 1 \leq i \leq k$.

(ix) Find h_i and h_{ijl}, $1 \leq i,\ j,\ l \leq k$, of which $h_{ijl} = h_{ilj}$, such that the theoretical moment $E\left(u_i^{k_1} u_j^{k_2}\right)$ of

$$u_i = h_i u_i^* + \sum_{j=1}^{k}\sum_{l=1}^{k} h_{ijl} u_j^* u_l^* - \sum_{j=1}^{k} h_{ijj}, \quad 1 \leq i \leq k,$$

computed by restricting the variance of u_i^* to one and using the approximate distribution QN$(0,\ \lambda^{(i)})$ for u_i^*, is approximately equal to $\hat{M}_{ij}^{(u)(k_1,\,k_2)}$, $1 \leq i,\ j \leq k,\ k_1 \geq 0,\ k_2 \geq 0,\ 2 \leq k_1 + k_2 \leq 4$.

(x) We may describe the distribution of $\mathbf{t}$ via the equation

$$\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{bmatrix} = \begin{bmatrix} \bar{t}_1 \\ \bar{t}_2 \\ \vdots \\ \bar{t}_k \end{bmatrix} + \mathbf{V} \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \vdots \\ \tilde{u}_k \end{bmatrix}$$

where $\tilde{u}_i = \sqrt{\operatorname{Var}(\tilde{u}_i)}u_i$, $u_i = h_i u_i^* + \sum_{j=1}^{k}\sum_{l=1}^{k} h_{ijl} u_j^* u_l^* - \sum_{j=1}^{k} h_{ijj}$

and

$$y = \begin{cases} \lambda_1^{(i)} z_i + \lambda_2^{(i)}\left(z_i^2 - \frac{1+\lambda_3^{(i)}}{2}\right), & z \geq 0, \\ \lambda_1^{(i)} z_i + \lambda_2^{(i)}\left(\lambda_3^{(i)} z_i^2 - \frac{1+\lambda_3^{(i)}}{2}\right), & z < 0. \end{cases}$$

Thus the equations in (10) provide a method for transforming $\mathbf{t}$ to $(z_1,\ z_2, \ldots,\ z_k)$ which is a set of uncorrelated standard normal random variables.

5 Comparison of powers of tests

Consider the case when the sample size is $n = 30$ and there are no outliers. A total of $N_s = 10{,}000$ values of y are generated from a normal distribution. Let the value of $(S,\ K)$ based on the n_s-th generated of y be denoted by $(t_{1n_s},\ t_{2n_s})$. By applying the method in Section 4, we can transform $(t_1,\ t_2)$ to $(z_1,\ z_2)$ which is a set of two uncorrelated standard normal variables.

We may denote the case when there are m_1 outliers at the lower end and m_2 outliers at the upper end by "m_1L m_2U". When $(m_1,\ m_2) = (0,\ 1),\ (0,\ 2)$ or $(1,\ 1)$, we can likewise generate $N_s = 10{,}000$ values of y from a normal distribution, compute $(t_{1n_s},\ t_{2n_s}) = (S*,\ K*)$ and transform $(t_1,\ t_2)$ to $(z_1,\ z_2)$. The values of the parameters of the distribution of $(S^*,\ K^*)$ are shown in Table 5.1.

Table 1: Parameters of the Distribution of $(S^*,\ K^*)$.

Parameter	OLOU	OL1U	OL2U	1L1U
$\bar{t}_1$	-4.51E-04	-0.18787	-0.29245	-5.64E-05
$\bar{t}_2$	2.809204	2.662833	2.648848	2.410794
V_{11}	-1	-0.943200	-0.933380	-0.999960
V_{12}	9.82E-04	-0.332220	-0.358890	0.008961
V_{21}	9.82E-04	-0.332220	-0.358890	0.008961
V_{22}	1	0.943202	0.933380	0.999960
h_1	0.998284	0.998474	0.990946	0.997411
h_2	0.760661	0.989107	0.995806	0.797483
$[\text{Var}(\tilde{u}_1)]^{1/2}$	0.405923	0.323826	0.282934	0.345384
$[\text{Var}(\tilde{u}_2)]^{1/2}$	0.694285	0.704974	0.802194	0.485362
h_{111}	0	0	0	0
h_{112}	0.01	-3.47E-18	-0.02	-3.47E-18
h_{122}	-0.02	-0.02	-0.04	-0.03
h_{211}	0.41	0.05	0.02	0.4
h_{212}	3.47E-18	0.05	0.04	-0.01
h_{222}	0	0	0	0
$\lambda_1^{(1)}$	0.885831	0.684336	0.691940	0.935871
$\lambda_2^{(1)}$	0.071763	0.014542	0.014289	0.044736
$\lambda_3^{(1)}$	-0.964280	-20.726100	-22.204300	-0.776450
$\lambda_1^{(2)}$	0.540055	0.656084	0.449040	0.697030
$\lambda_2^{(2)}$	0.454883	0.424341	0.548370	0.353401
$\lambda_3^{(2)}$	-0.105370	0.209001	-0.048130	0.083958

We may use the rejection region

$$\{(S*,\ K*):\ z_1^2 + z_2^2 > \chi^2_{2,0.05} = 5.99\}$$

for testing the hypothesis of normality at the 0.05 level.

The power of the JB test with rejection region $\{(S,\ K)\ :\ \text{JB} > 4.1797\}$, and those of the test based on a circle in the $z_1 - z_2$ plane derived from $(S^*,\ K^*)$ are shown in Table 5.2. The columns labeled by $\bar{m}_3$ and $\bar{m}_4$ give respectively the coefficients of skewness and kurtosis of the variable y_i.

When $(\bar{m}_3,\ \bar{m}_4) = (0,\ 3)$ in which case the y_i are normally distributed, Table 5.2 shows that the powers of the JB test and the test based on a circle in the $z_1 - z_2$ plane are all not too far from the target value 0.05.

The columns labeled by JB and OL0U in Table 5.2 show that in rows 6-22, the powers in the OL0U column are larger than those in the JB column by more than 2%. Thus in the case when there are no outliers, the test based on a circle in the $z_1 - z_2$ plane is a strong competitor to the JB test.

Table 2: Powers of JB test and test based on a circle in the $z_1 - z_2$ plane (n = 30, α =0.05, N_s = 10,000).

No.	$\bar{m}_3$	$\bar{m}_4$	JB	OL0U	OL1U	OL2U	1L1U
1	0	2.2	0.002750	0.013950	0.011100	0.019700	0.026500
2	0	2.8	0.030800	0.042050	0.034900	0.041000	0.048400
3	0	3	0.053850	0.068800	0.058600	0.055700	0.063800
4	-0.1	3.2	0.079400	0.098200	0.080900	0.073000	0.084000
5	0.1	3.2	0.080250	0.095850	0.077300	0.072000	0.076900
6	0	4	0.187050	0.220900	0.184200	0.151300	0.172100
7	0.3	4	0.188750	0.216100	0.178300	0.143800	0.163800
8	-0.3	4	0.194000	0.219400	0.185100	0.175900	0.169400
9	-0.1	5	0.311500	0.361850	0.294500	0.270800	0.288900
10	0.5	5	0.312100	0.349800	0.284800	0.236900	0.284500
11	0.1	5	0.314700	0.369550	0.297100	0.264000	0.295200
12	-0.5	5	0.316100	0.349700	0.304500	0.282200	0.282500
13	0	6	0.426450	0.490500	0.405000	0.370500	0.410600
14	1	10	0.685700	0.749050	0.679600	0.615700	0.698100
15	-1	10	0.687350	0.746850	0.671300	0.665000	0.710400
16	0.1	10	0.699250	0.770750	0.706600	0.656100	0.738000
17	0	10	0.701400	0.772700	0.703300	0.656100	0.743200
18	-0.1	10	0.704100	0.773550	0.705500	0.662200	0.747600
19	-1.5	11	0.717050	0.765100	0.709300	0.701100	0.720400
20	1.5	11	0.718150	0.762350	0.714600	0.646200	0.725800
21	2	15	0.834400	0.875700	0.852300	0.800300	0.873400
22	-2	15	0.838350	0.875800	0.835800	0.832000	0.869100
23	-2.5	14	0.856050	0.868300	0.801000	0.800900	0.832300
24	2.5	14	0.856750	0.869400	0.839800	0.781400	0.828600
25	3	19	0.926700	0.935500	0.920600	0.887800	0.922200
26	-3	17	0.956100	0.971300	0.901000	0.898200	0.955300
27	3.5	22	0.977450	0.982350	0.979400	0.970900	0.984600
28	-3.5	22	0.978100	0.983950	0.965400	0.964000	0.984500
29	3.8	24.5144	0.997500	0.999000	0.997700	0.996900	0.998000
30	-3.8	24.4	0.998550	0.999450	0.990400	0.987100	0.999700

The columns labeled by OL0U and OL1U in Table 5.2 show that when there is an outlier at the upper end, the test based on a circle in the $z_1 - z_2$ plane suffers a slight loss in the power of the test. The columns labeled by OL1U and OL2U show that when the number of outliers at the upper end increases by one, the power of the test tends to decrease further. The columns labeled by OLOU and 1L1U also show that there is a slight loss in the power of the test when there are outliers on both ends.

6 Concluding Remarks

The test based on a circle with centre zero in the $z_1 - z_2$ plane tends to have good power. This is not surprising because the circle with centre zero is the smallest region in the $z_1 - z_2$ plane with the given probability. When the number of outliers is small, there is a slight loss in the power of the test. When the number of outliers is large, it is likely that the loss in power would be large.

References

[1] Jargue, C. M. and A. K. Bera, 1980. "Efficient tests for normality, homoscedasticity and serial dependence of regression residuals," *Econ. Lett.*, **6**: 255-259.

[2] Shapiro, S. S. and Wilk, M. B., 1965. "An Analysis of variance test for normality (complete samples)," *Biometrika*, **52 (3-4)**: 591-611.

[3] Anderson, T. W. and Darling, D. A., 1952. "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes," *Annals of Mathematical Statistics*, **23**: 193-212.

[4] Gel, Y. R. and J.L. Gastwirth, 2007. "A robust modification of the Jargue-Bera test of normality". *Econ. Lett.*, **99**: 30-32.

[5] Imon, A.H.M.R., 2003. Regression residuals, moments and their use in tests for normality. *Commun. Stat. Theor. Methods.*, **32**: 1021-1034.

[6] Rana, Midi and Imon, 2009. "A Robust Rescaled Moment Test for Normality in Regression," *Journal of Math and Stat.*, **5(1)**: 54-62.

[7] Pooi A. H., 2003. *Effects of non-normality on confidence intervals in linear models.* Technical Report No.6/2003. Institute of Mathematical Sciences, University of Malaya.

[8] Pooi A. H., 2006. *Non-normality distributed variates with a nonlinear dependence structure.* Technical Report No.13/2006. Institute of Mathematical Sciences, University of Malaya.