Journal of Applied Statistics
Vol. 34, No. 1, 87–105, January 2007

Routledge
Taylor & Francis Group

Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison

THORSTEN THADEWALD & HERBERT BÜNING

Institute for Statistics and Econometrics, Free University Berlin, Germany

ABSTRACT *For testing normality we investigate the power of several tests, first of all, the well-known test of Jarque & Bera (1980) and furthermore the tests of Kuiper (1960) and Shapiro & Wilk (1965) as well as tests of Kolmogorov–Smirnov and Cramér-von Mises type. The tests on normality are based, first, on independent random variables (model I) and, second, on the residuals in the classical linear regression (model II). We investigate the exact critical values of the Jarque–Bera test and the Kolmogorov–Smirnov and Cramér-von Mises tests, in the latter case for the original and standardized observations where the unknown parameters* μ *and* σ *have to be estimated. The power comparison is carried out via Monte Carlo simulation assuming the model of contaminated normal distributions with varying parameters* μ *and* σ *and different proportions of contamination. It turns out that for the Jarque–Bera test the approximation of critical values by the chi-square distribution does not work very well. The test is superior in power to its competitors for symmetric distributions with medium up to long tails and for slightly skewed distributions with long tails. The power of the Jarque–Bera test is poor for distributions with short tails, especially if the shape is bimodal – sometimes the test is even biased. In this case a modification of the Cramér-von Mises test or the Shapiro–Wilk test may be recommended.*

KEY WORDS: Goodness-of-fit tests, tests of Kolmogorov–Smirnov and Cramér-von Mises type, Shapiro–Wilk test, Kuiper test, skewness, kurtosis, contaminated normal distribution, Monte Carlo simulation, critical values, power comparison

Introduction

Goodness-of-fit tests play an important role in statistical applications, especially in the case of testing univariate normality, see for example D'Agostino & Stephens (1986). Normality may be the most common assumption in applying statistical procedures as in the classical linear regression model where the (unobserved) disturbance vector ε is assumed to be normally distributed. It is well known that departures from normality may lead to substantially incorrect statements in the analysis of economic models. Thus, a test on normality based on the (observable) regression residuals is an absolute 'must' in any regression analysis. Here, we restrict our attention to the vector $\hat{\varepsilon}$ of OLS

Correspondence Address: Herbert Büning, Institute for Statistics and Econometrics, Free University Berlin, Boltzmannstr. 20, 14195 Berlin, Germany. Email: herbert.buening@wiwiss.fu-berlin.de

0266-4763 Print/1360-0532 Online/07/010087–19
DOI: 10.1080/02664760600994539

residuals, which is given by $\hat{\varepsilon} = (I - H)\varepsilon$ with $H = X(X'X)^{-1}X'$, i.e. the vector of residuals $\hat{\varepsilon}$ is a linear transformation of the unobserved disturbance vector ε. Therefore, tests of distributional assumptions on ε are based on $\hat{\varepsilon}$.

One of the most famous tests for normality of regression residuals is the test of Jarque & Bera (1980, 1987), which has gained great acceptance among econometricians. The test statistic *JB* is a function of the measures of skewness S and kurtosis K computed from the sample. Under normality, the theoretical values of S and K are 0 and 3, respectively. The purpose of this paper is to compare the Jarque–Bera test with other goodness-of-fit tests like the Shapiro–Wilk test, the Kuiper test as well as with tests of Kolmogorov–Smirnov and Cramér-von Mises type in varying sample situations. As pointed out by several authors, see for example Jarque & Bera (1987) and Urzua (1996), the Jarque–Bera test behaves well in comparison with some other tests for normality discussed in the literature if the alternatives to the normal distribution belong to the Pearson family. For our power study we assume the model of contaminated normal distributions (CN) for the components ε_i of disturbance vector ε, i.e. $\varepsilon_i \sim F = (1-p)N(\mu_1, \sigma_1^2) + pN(\mu_2, \sigma_2^2)$, $i = 1, \ldots, n$, $0 \leq p \leq 1$. This model covers a broad range of distributions, symmetric and asymmetric ones, and can also be used to describe ‘small departures from normality.

The contaminated normal distribution function F can be interpreted as follows: with probability $(1 - p)$, an observation comes from a normal distribution with mean μ_1 and variance σ_1^2 and with probability p from a normal distribution with mean μ_2 and variance σ_2^2. Notice, that F is not the c.d.f. of a normal distribution. In the next section we describe the two models and the hypotheses and give two data examples, one concerning the first model (random sample) and one the second model (linear regression). The third section presents all the tests mentioned above and the fourth section deals with a power comparison of the tests including a study on the critical values of the Jarque–Bera test and the tests of Kolmogorov–Smirnov and Cramér-von Mises type. The power comparison is carried out via Monte Carlo simulation. As a result, the Jarque–Bera test is, on the whole, the best one among all tests considered but for special sample situations it does not work very well and other tests should be preferred. The fifth section gives a resumé of the results and some hints for further studies.

Model, Hypotheses and Data Examples

We consider two models, a random sample and the classical linear regression.

Model I (Random Sample)

Let be $X_1, \ldots, X_n$ independent random variables with absolutely continuous distributions function F. We wish to test

$$H_0{:}F(x) = \Phi((x-\mu)/\sigma) \quad \text{for all} \quad x \in R$$

versus the two-sided alternative $H_1{:}F(x) \neq \Phi((x-\mu)/\sigma)$ for at least one $x \in R$, where Φ is the c.d.f. of the standard normal distribution and μ $(-\infty < \mu < \infty)$ as well as σ $(\sigma > 0)$ may be known or unknown. In the case of known μ and σ we assume without any loss of generality $\mu = 0$ and $\sigma = 1$. In the case of unknown μ and σ the parameters are estimated by the sample mean $\bar{x}$ and sample standard deviation s, respectively.

Model II (Classical Linear Regression)

Let $y = X\beta + \varepsilon$, where y is a $(n, 1)$-vector, X is a non-stochastic (n, k)-matrix of rank k, β is the $(k, 1)$-vector of unknown parameters and ε is the disturbance $(n, 1)$-vector whose components are assumed to be uncorrelated and distributed with expectation zero and constant variance σ^2.

Since the vector ε is unobservable, a test for normality generally is based on sample residuals such as OLS residuals $\hat{\varepsilon}_i$ which are given by $\hat{\varepsilon} = (I - H)\ \varepsilon$ with the so-called hat matrix $H = X(X'X)^{-1}X'$. Note, that the components $\hat{\varepsilon}_i$ of $\hat{\varepsilon}$ are not uncorrelated and do not have equal variances in general.

Now, let us present two real data examples to which the selected tests are applied later on, example 1 for model I and example 2 for model II.

Example 1

The first example refers to returns of the share index DAX. The DAX measures the performance of the Prime Standards 30 largest German companies. The data set contains 200 daily DAX returns (in percent) from June 18, 2002, until March 31, 2003.

Example 2

The dependency of corporate revenues on advertising spending of leading US advertisers is the subject of the second example. The following data set shows the total US ad spending and US corporate revenue (in million dollars for 2001) of 50 large companies, e.g. General Motors, Ford, Daimler-Chrysler, AT&T, IBM, Hewlett-Packard, Sony, Walt Disney, Coca-Cola, PepsiCo and Unilever.

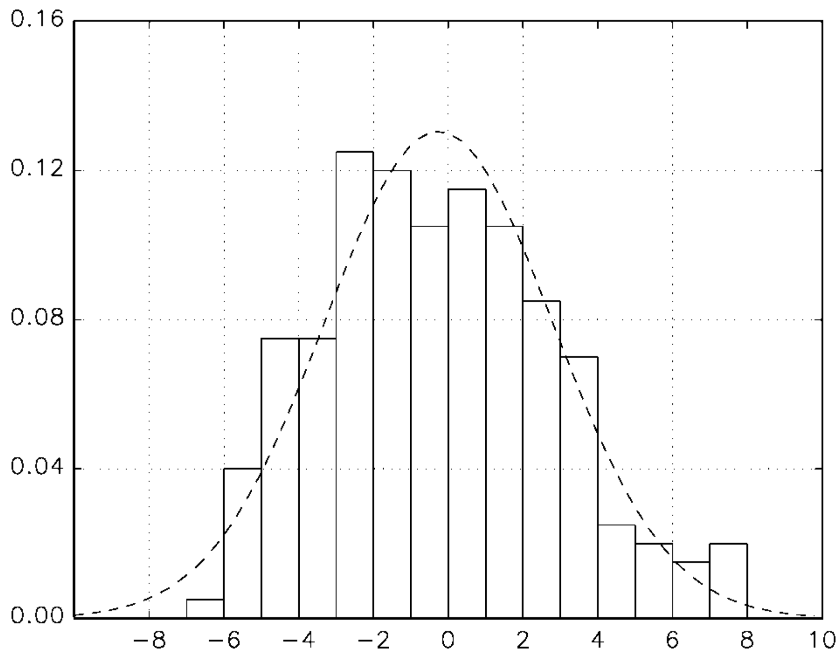

Figure 1. Histogram of daily DAX returns

Table 1. Data set U.S. advertisers 2001

Ad spend.	Corp. rev.	Ad spend.	Corp. rev.	Ad spend.	Corp. rev.
430	2354	289	5528	449	3414
444	17522	664	74476	1137	39900
656	12356	884	6745	385	16726
1372	52550	3374	132399	552	7716
421	5804	881	14525	290	9452
339	37668	366	4557	2210	18215
543	19597	899	31725	479	3534
974	13154	778	53553	812	10131
397	5021	1103	34673	1310	21127
527	7842	994	35215	926	39888
410	12791	426	9382	1484	11315
903	7526	1086	32004	1462	64649
1985	72708	1618	20204	1283	19466
302	21760	422	6129	596	15980
289	3997	304	9328	1757	20970
746	15651	512	50098	312	1724
2408	108296	428	9363		

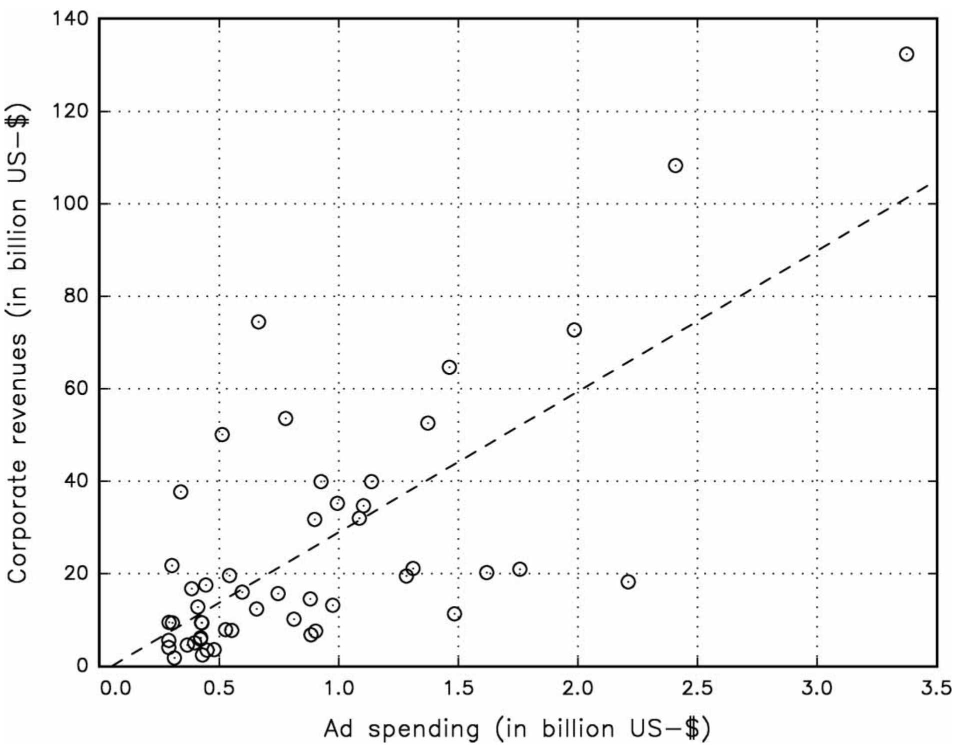

Figure 2. Scatterplot of U.S. advertisers 2001

Goodness-of-Fit Tests

Jarque–Bera Test and its Modification

The test statistic JB of Jarque–Bera is defined by

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

where the sample skewness $S = \hat{\mu}_3 / \hat{\mu}_2^{3/2}$ is an estimator of $\beta_1 = \mu_3 / \mu_2^{3/2}$ and the sample kurtosis $K = \hat{\mu}_4 / \hat{\mu}_2^2$ an estimator of $\beta_2 = \mu_4 / \mu_2^2$, μ_2 and μ_3 are the theoretical second, third and fourth central moments, respectively, with its estimates

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^j, \quad j = 2, 3, 4$$

JB is asymptotically chi-squared distributed with two degrees of freedom because *JB* is just the sum of squares of two asymptotically independent standardized normals, see Bowman & Shenton (1975). That means: H_0 has to be rejected at level α if $JB \geq \chi^2_{1-\alpha,\, 2}$.

In the more usual case of linear regression *JB* is calculated for the regression residuals.

Urzúa (1996) introduced a modification of the Jarque–Bera test – we call it JBU – by standardizing the skewness S and kurtosis K in the formula of JB appropriately in the following way:

$$\mathrm{JBU} = \left(\frac{S^2}{v_S} + \frac{(K - e_K)^2}{v_K} \right)$$

with

$$v_S = \frac{6(n-2)(n+1)}{(n+1)(n+3)}$$

$$e_K = \frac{3(n-1)}{n+1}$$

and

$$v_K = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

Notice, that JB and JBU are asymptotically equivalent, i.e. H_0 has to be rejected at level α if $\mathrm{JB} \geq \chi^2_{1-\alpha,\, 2}$.

Shapiro–Wilk Test

The Shapiro–Wilk test is motivated by a probability plot in which we consider the regression of the ordered observations on the expected values of the order statistics from the hypothesized distribution. The test statistic is defined by

$$\mathrm{SW} = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)} \right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$X = (X_1, \ldots, X_n)$ is a vector of random variables and $X_{()}$ the corresponding ordered vector. $\bar{X}$ is the usual sample mean. The weights a_i, $i = 1, \ldots, n$, are calculated like this. $Y = (Y_1, \ldots, Y_n)$ is a vector of random variables from a normal distribution and $Y_{()}$ again the corresponding ordered vector. The determination of a_i requires the calculation of the vector of expectation values and the covariance matrix of $Y_{()}$: $m' = (m_1, \ldots, m_n)$

where $m_i = \mathrm{E}(Y_{(i)})$ and V where $v_{ij} = \mathrm{Cov}(Y_{(i)}, Y_{(j)})$. The vector a of the weights a_i yields as follows: $a' = m'V^{-1}[(m'V^{-1})(V^{-1}m)]^{-1/2}$. H_0 has to be rejected, if $\mathrm{SW} \leq w_\alpha$.

For the components of the vector a we have $a_i = -a_{n-i+1}$, they are tabulated by Shapiro & Wilks (1965) for $n \leq 50$, where critical values w_α of SW are given, too, see also Shapiro *et al*. (1968) and Shapiro & Francia (1972).

Tests of Kolmogorov–Smirnov Type

Now, let us consider non-parametric goodness-of-fit tests that are based on the empirical distribution function.

1. Kolmogorov–Smirnov test KS
   As above, let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of $X_1, \ldots, X_n$ and let F_n be the usual empirical distribution functions for the X-sample. Then the Kolmogorov–Smirnov statistic KS is defined by

$$\mathrm{KS} = \sup_x |F_n(x) - F_0(x)|$$

   In our case of testing normality we have $F_0(x) = \Phi((x - \mu)/\sigma)$.
   The corresponding test rejects H_0 if $\mathrm{KS} \geq k_{1-\alpha}$. Exact critical values $k_{1-\alpha}$ are reported by Büning & Trenkler (1994), where the asymptotic null distribution of KS is given, too.

2. Modified KS-test
   The test statistic KS can be modified by introducing appropriate non-negative weight functions $W(F_0(x))$ in order to give different weights to the difference $|F_n(x) - F_0(x)|$, see Büning (2001). As a special case we choose $W(F_0(x)) = \sqrt{F_0(x)(1 - F_0(x))/n}$ which is symmetric about the centre $F_0(x) = 0.5$. This weight function was introduced by Anderson & Darling (1952).
   Then we can define a modification of KS by

$$\mathrm{KSW} = \sup_x \frac{\sqrt{n}|F_n(x) - F_0(x)|}{\sqrt{F_0(x)(1 - F_0(x))}}.$$

   Obviously, the denominator $\sqrt{F_0(x)(1 - F_0(x))}$ places a high weight on the upper and lower part of the underlying distribution.
   The corresponding test rejects H_0 if $\mathrm{KSW} \geq k^{(w)}_{1-\alpha}$. Critical values $k^{(w)}_{1-\alpha}$ are reported in Table 2.

3. Kuiper test
   Let be $D^+ = \sup_x (F_n(x) - F_0(x))$ and $D^- = \sup_x (F_0(x) - F_n(x))$. Then the statistic of Kuiper (1960) is defined by

$$\mathrm{KUI} = D^+ + D^-.$$

   It should be noted that the Kolmogorov–Smirnov statistic KS can be written as $\mathrm{KS} = \max(D^+, D^-)$.
   H_0 has to be rejected if $\mathrm{KUI} \geq k^{(u)}_{1-\alpha}$. Critical values $k^{(u)}_{1-\alpha}$ are given in Table 2.

Tests of Cramér-von Mises Type

1. Cramér-von Mises test
   The Cramér-von Mises statistic CM is defined as follows:

$$\mathrm{CM} = n \cdot \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 f_0(x)dx,$$

   which can be written as

$$\mathrm{CM} = \frac{1}{12n} + \sum_{i=1}^{n}\left[F_0(x_{(i)}) - \frac{2i-1}{2n}\right]^2.$$

   The test rejects H_0 if $\mathrm{CM} \geq c_{1-\alpha}$. Approximate critical values $c_{1-\alpha}$ can be found in Anderson & Darling (1952).

2. Modified CM-test
   In the same manner as for the KS-test we can modify the CM-test by introducing an appropriate weight function. Here, we choose the weight function $W(F_0(x)) = F_0(x)(1 - F_0(x))$, i.e.

$$\mathrm{CMW} = n \cdot \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} f_0(x)dx \quad \text{for} \quad F_0(x) \neq 0, 1.$$

   For computational simplification CMW can be written as

$$\mathrm{CMW} = n - \frac{1}{n}\sum_{i=1}^{n} (2i-1)(\ln p_{(i)} + \ln(1 - p_{(n+1-i)})) \quad \text{with} \quad p_{(i)} = F_0(x_{(i)}).$$

   The corresponding test rejects H_0 if $\mathrm{CMW} \geq c^{(w)}_{1-\alpha}$. Critical values $c^{(w)}_{1-\alpha}$ are reported in Table 2.

In Table 2, critical values of the test statistics KSW, CMW and KUI are presented for selected sample sizes and $\alpha = 0.05$. The critical values were obtained by Monte Carlo simulation (100.000 replications). The suffix '_S indicates a test based on standardized observations. For example, KSS is the abbreviation for the corresponding KS test. Critical values of the KSW-test can also be found, for example, in Canner (1975).

For a comprehensive study of tests based on the empirical distribution function, see Stephens (1974).

Applications

Now, let us test the hypothesis of normality for the data in Examples 1 and 2 by applying all the tests above. We get the results given in Table 3 and Table 4, respectively.

Obviously, the results for the eight tests are extremely different. In Example 1, the Kuiper test has the greatest p-value of 49% and the Shapiro–Wilk test the smallest one

Table 2. Critical values of tests for various sample sizes n, $\alpha = 0.05$

Tests	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
JB	2.5347	3.7677	5.0037	5.4479	5.7275	5.8246
JBU	7.4374	6.8559	6.5940	6.3381	6.1678	6.0378
KS	0.4101	0.2936	0.1886	0.1338	0.0951	0.0603
KSS	0.2615	0.1920	0.1243	0.0890	0.0635	0.0403
KSW	2.0236	1.4369	0.9165	0.6451	0.4544	0.2869
KSWS	0.8535	0.8049	0.6637	0.5266	0.4007	0.2697
KUI	0.5142	0.3719	0.2401	0.1715	0.1217	0.0774
KUIS	0.4318	0.3166	0.2049	0.1469	0.1046	0.0665
CM	0.4584	0.4534	0.4601	0.4567	0.4627	0.4598
CMS	0.1193	0.1230	0.1241	0.1253	0.1258	0.1265
CMW	2.5265	2.4737	2.4868	2.4695	2.5280	2.4814
CMWS	0.6905	0.7176	0.7380	0.7460	0.7479	0.7506
SW	0.8451	0.9040	0.9486	0.9642	0.9727	0.9795

Table 3. Testing normality for DAX data (short tails)

Results of testing normality (Example 1: daily DAX returns)								
Test	JB	JBU	KS	KSW	KUIP	CM	CMW	SW
Statistic	5.1218*	5.1388	0.0460	0.2050	0.0751	0.0801	0.6071	0.9827
p-Wert	0.062	0.072	0.38	0.22	0.49	0.21	0.22	0.015
Asymp. p	0.077							

*$S = 0.3433$ and $K = 2.6216$.

Table 4. Testing normality for advertising data (residuals)

Results of testing normality (Example 2: residuals of U.S. corporate revenues on U.S. ad spending)								
Test	JB	JBU	KS	KSW	KUIP	CM	CMW	SW
Statistic	3.9716*	5.6718	0.1386	0.4726	0.2257	0.1666	0.9099	0.9594
p-Wert	0.071	0.062	0.017	0.115	0.015	0.014	0.019	0.084
Asymp. p	0.137							

*$S = 0.4600$ and $K = 4.0296$.

with 1.5%, it is the only test that rejects at the 5% level. The differences of the p-values in Example 2 are not so considerable as in the first example, but now, half of the tests reject at the 5% level. Thus, a power comparison of all the tests becomes important, it is carried out in the next section.

Power study

Critical Values of Some Tests

Exact critical values of JB for models I and II. For model I and II let us compare the asymptotic critical values of the Jarque–Bera test JB with the exact ones calculated by simulation for sample sizes $n = 10$, 20, 50, 100, 200 and 500 and varying levels 0.01, 0.02, 0.05, 0.1 and 0.2. In model II we consider the cases of one, three and six independent regression variables assuming different distribution functions such as uniform, normal and exponential. Table 5 displays the critical values of JB for model I.

From Table 5 we can state that the JB-test based on asymptotical critical values is conservative for $\alpha \geq 0.05$, and that considerably so for small sample sizes; for $\alpha < 0.05$, the pattern is not so clear. As an example: for $\alpha = 0.01$, the approximation by the Chi-square distribution is better for $n = 20$ than for $n = 200$, a surprising result. In general, the approximation by the Chi-squared distribution does not work well even not for large sample sizes, the speed of convergence is very slow. Thus, approximate critical values $\chi^2_{1-\alpha}$ for the JB-test should be used cautiously in empirical studies and for a meaningful power comparison exact critical values have to be used.

Figure 3 presents the graphs of the exact and asymptotic distribution of the JB-statistic. Obviously, the approximation becomes more precise with increasing sample sizes.

Now, let us consider model II. Table 6 lists critical values of JB with three independent regression variables from a standard normal distribution. We restrict our presentation to this case because our simulation has shown that the pattern of critical values is nearly the same for one, three and six variables as well as for other distributions mentioned above.

From Tables 5 and 6 we see that for all levels the differences between the critical values in model I and II are negligible for $n \geq 50$, the pattern in model II is nearly the same as in model I.

Exact critical values of KS and CM for original and standardized data. We study critical values of the KS-test and CM-test for the original data and for the standardized observations where the unknown parameters μ and σ of the normal distribution function are estimated by the arithmetic mean $\bar{x}$ and the empirical standard deviation s_x, i.e. $z_i = (x_i - \bar{x})/s_x$. We again consider sample sizes $n = 10$, 20, 50, 100, 200 and 500, but here only the usual levels $\alpha = 0.01$, 0.05 and 0.10.

From Table 7 we can state that for the tests KS and CM there is a great difference between the critical values of original and standardized data. If we have to estimate the unknown parameters μ and σ and apply, however, both tests by using critical values of KS and CM based on original data then the tests are extremely conservative. The same is true for the modifications KSW and CMW of KS and CM, respectively.

Table 5. Critical values of JB-test (random sample)

Critical values of Jarque–Bera test (random sample)							
α	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n \to \infty$
0.01	5.738	9.458	12.331	12.296	11.750	10.601	9.210
0.02	4.274	6.583	8.721	9.089	8.788	8.349	7.824
0.05	2.535	3.768	5.004	5.448	5.728	5.825	5.991
0.10	1.618	2.335	3.192	3.643	4.081	4.324	4.605
0.20	1.126	1.556	2.122	2.474	2.748	2.985	3.219

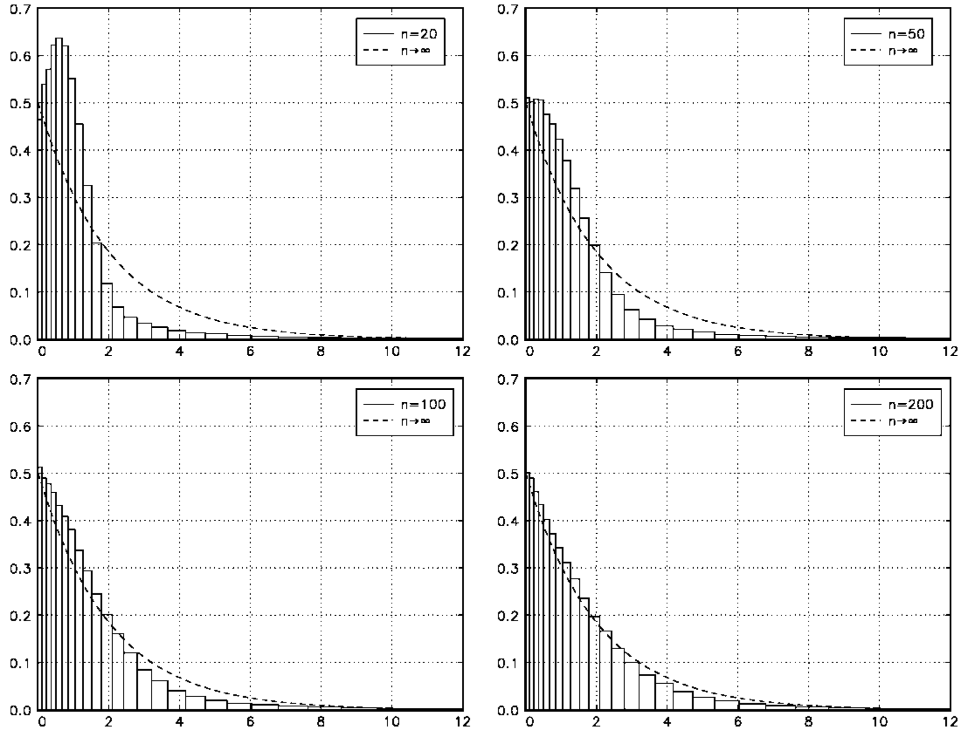

Figure 3. Simulated and asymptotic distribution of the JB-test in comparison

Figure 4 illustrates the great difference between the distributions of KS, CM (original data) and KSS, CMS (standardized data).

Power Comparison

We investigate via Monte Carlo simulation (10.000 replications) the power of all the tests presented in the third section. To conduct the simulation study we select the model of contaminated normal CN for H_1, the distribution function F of which can be given in the following form, see the first section: $F = (1-p)N(\mu_1, \sigma_1^2) + pN(\mu_2, \sigma_2^2)$ with $0 \leq p \leq 1$.

Without any loss of generality we assume $\mu_1 = 0$ and $\sigma_1^2 = 1$. Furthermore, we choose $\mu_2 = 0, 1, 2, 3, 4$, $\sigma_2 = 1, 2, 3, 4, 6$ and $p = 0.01, 0.05, 0.20, 0.35, 0.50, 0.65, 0.80, 0.90,$

Table 6. Critical values of JB-test (residuals)

Critical values of Jarque–Bera test (residuals of a regression with three independent variables)						
α	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
0.01	4.8219	9.6598	12.1265	12.2825	11.7110	10.6732
0.02	3.7038	6.6847	8.4769	8.8522	8.7956	8.4148
0.05	2.3291	3.8649	4.9556	5.4183	5.6596	5.8275
0.10	1.4789	2.3681	3.1673	3.6801	4.0355	4.3096
0.20	1.0713	1.5682	2.1243	2.4857	2.7626	2.9967

Table 7. Critical values of KS- and CM-test

α	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
Critical values of KS (original data)						
0.01	0.4898	0.3494	0.2248	0.1600	0.1142	0.0725
0.05	0.4101	0.2936	0.1886	0.1338	0.0951	0.0603
0.10	0.3701	0.2640	0.1699	0.1204	0.0857	0.0543
Critical values of KS (standardized data)						
0.01	0.3043	0.2224	0.1445	0.1035	0.0738	0.0470
0.05	0.2615	0.1920	0.1243	0.0890	0.0635	0.0403
0.10	0.2414	0.1761	0.1143	0.0817	0.0582	0.0369
Critical values of CM (original data)						
0.01	0.7143	0.7147	0.7341	0.7448	0.7376	0.7445
0.05	0.4584	0.4534	0.4601	0.4567	0.4627	0.4598
0.10	0.3469	0.3439	0.3487	0.3454	0.3494	0.3458
Critical values of CM (standardized data)						
0.01	0.1687	0.1733	0.1754	0.1767	0.1780	0.1779
0.05	0.1193	0.1230	0.1241	0.1253	0.1258	0.1265
0.10	0.0989	0.1015	0.1024	0.1034	0.1033	0.1036

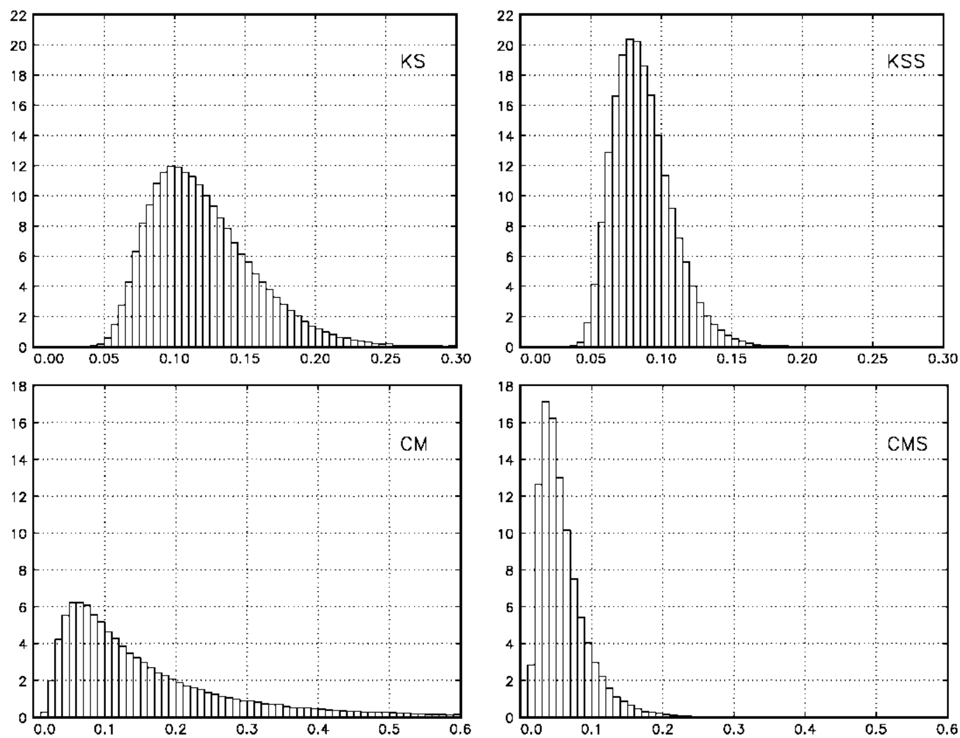

Figure 4. Distributions of KS, KSS, CM, CMS ($n = 50$)

0.95, 0.99. The choice of such parameters guarantees a broad range of distributions, unimodal and bimodal ones, symmetric distributions with short up to very long tails and asymmetric ones with different strength of skewness. For all the tests we use exact critical values. In the simulation the tests of Kolmogorov–Smirnov and Cramér-von Mises type are carried out on a basis of standardized observations.

At first, comparing the JB-test and JBU-test we can state that the difference in power of the tests is very small, in 93% of the cases the difference is less than 1%, the JB-test seems to be a little bit more efficient than JBU. The power of JBU is never higher than 2% in comparison to JB, but in some cases lower than 10%. Because of the small power differences between JB and JBU we omit JBU from the following power presentations.

Figures 5 to 12 display the power of the seven tests JB, KS, KSW, KUI, CM, CMW and SW from the third section by selecting different values of μ_2, σ_2 and p. Different sample sizes are chosen in order to produce graphs for a visible power comparison. In the bottom of the figures the corresponding density functions of CN are plotted in comparison to the standard normal density. At first, we consider the case of random sample with original data and then the case of regression variables.

The values of the corresponding parameters of skewness β_1 and kurtosis β_2 are given in Tables 8 and 9.

Figures 5 to 8 are concerned with power curves as functions of σ_2 for different values of p and $\mu_2 = 0$ (symmetric case) and $\mu_2 = 3$ (asymmetric case for $\sigma_2 \neq 1$). Of course, with

Table 8. Skewness and kurtosis for various CN distributions, part 1

			Skewness S and kurtosis K				
Fig.	μ_2	p	$\sigma_2 = 1$	$\sigma_2 = 2$	$\sigma_2 = 3$	$\sigma_2 = 4$	$\sigma_2 = 6$
5	0	0.10	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$
			$\beta_2 = 3$	$\beta_2 = 4.44$	$\beta_2 = 8.33$	$\beta_2 = 12.72$	$\beta_2 = 19.33$
6	0	0.50	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$
			$\beta_2 = 3$	$\beta_2 = 4.08$	$\beta_2 = 4.92$	$\beta_2 = 5.34$	$\beta_2 = 5.68$
7	0	0.80	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$
			$\beta_2 = 3$	$\beta_2 = 3.37$	$\beta_2 = 3.56$	$\beta_2 = 3.64$	$\beta_2 = 3.7$
8	3	0.50	$\beta_1 = 0$	$\beta_1 = 0.65$	$\beta_1 = 0.92$	$\beta_1 = 0.96$	$\beta_1 = 0.83$
			$\beta_2 = 2.04$	$\beta_2 = 2.85$	$\beta_2 = 3.72$	$\beta_2 = 4.37$	$\beta_2 = 5.11$

Table 9. Skewness and kurtosis for various CN distributions, part 2

			Skewness S and kurtosis K				
Fig.	σ_2	p	$\mu_2 = 0$	$\mu_2 = 1$	$\mu_2 = 2$	$\mu_2 = 3$	$\mu_2 = 4$
9	4	0.05	$\beta_1 = 0$	$\beta_1 = 0.9$	$\beta_1 = 1.71$	$\beta_1 = 2.35$	$\beta_1 = 2.84$
			$\beta_2 = 13.47$	$\beta_2 = 14.12$	$\beta_2 = 15.75$	$\beta_2 = 17.65$	$\beta_2 = 19.24$
10	1	0.50	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$
			$\beta_2 = 3$	$\beta_2 = 2.92$	$\beta_2 = 2.5$	$\beta_2 = 2.04$	$\beta_2 = 1.72$
Fig.	μ_2	σ_2	$p = 0.05$	$p = 0.20$	$p = 0.5$	$p = 0.8$	$p = 0.95$
11	0	4	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$
			$\beta_2 = 13.47$	$\beta_2 = 9.75$	$\beta_2 = 5.34$	$\beta_2 = 3.64$	$\beta_2 = 3.14$

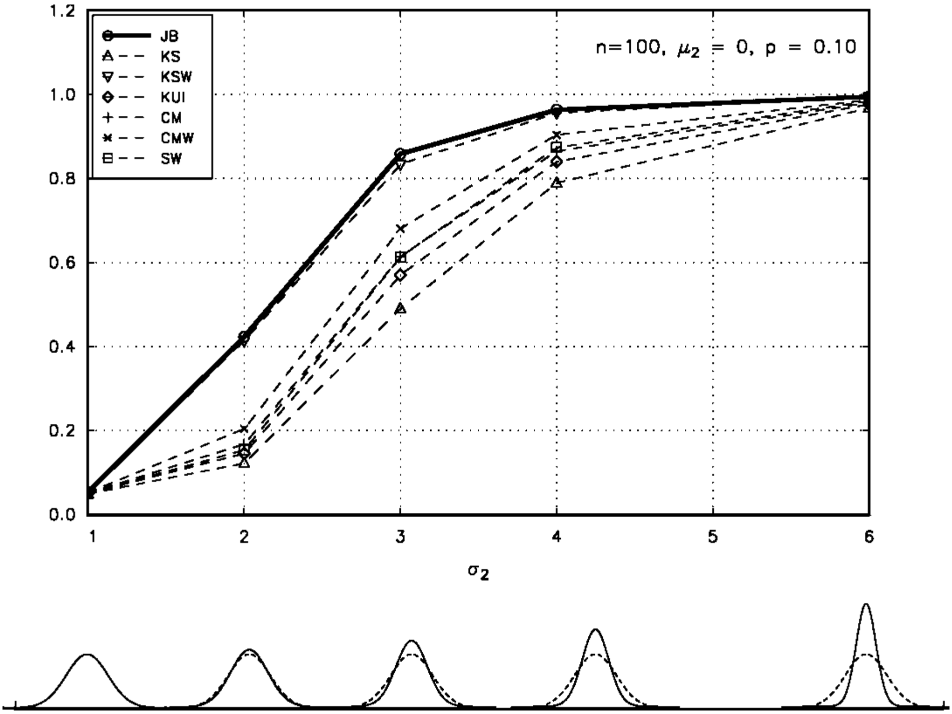

Figure 5. Power of the tests (symmetric case 1, random sample)

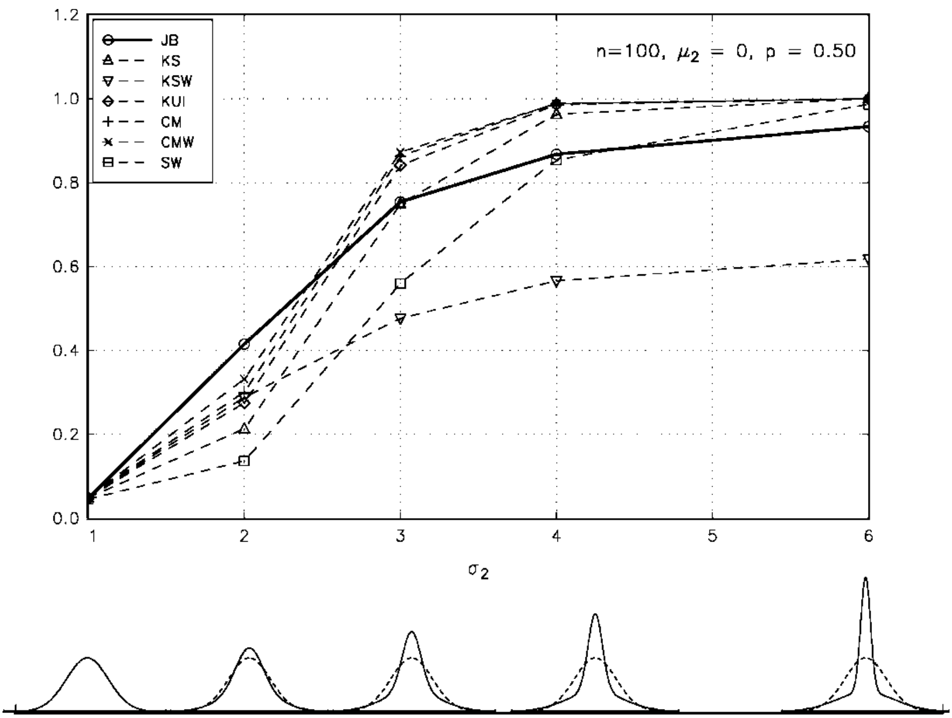

Figure 6. Power of the tests (symmetric case 2, random sample)

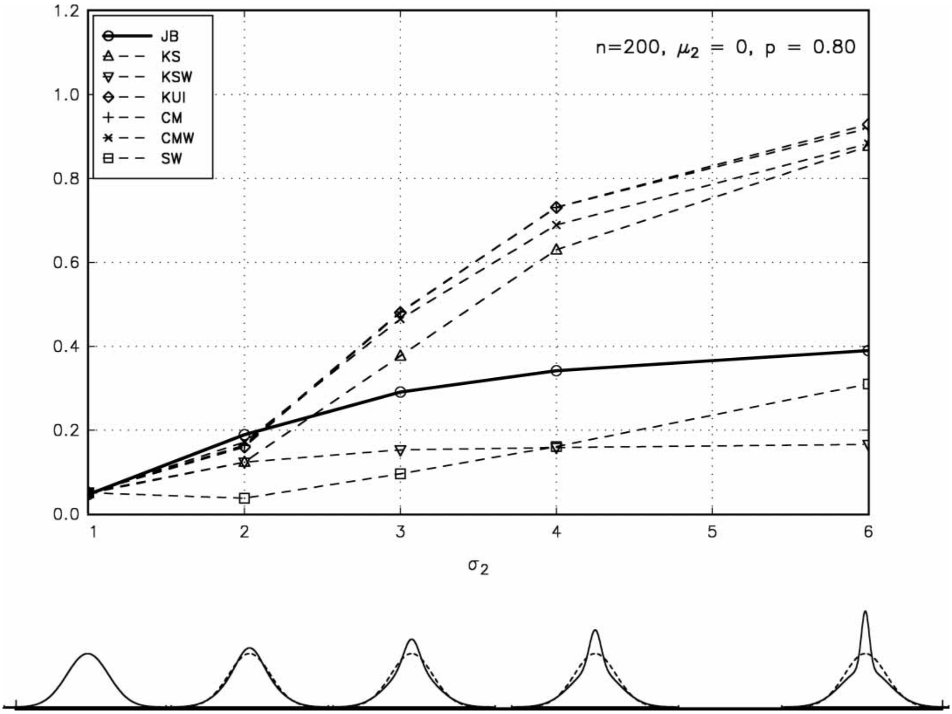

Figure 7. Power of the tests (symmetric case 3, random sample)

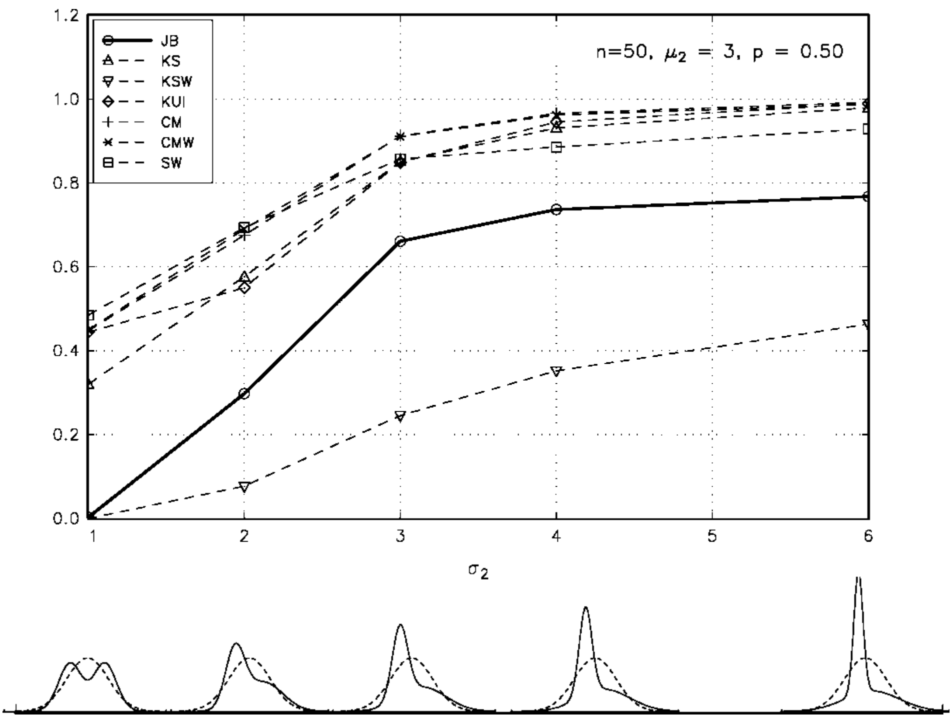

Figure 8. Power of the tests (asymmetric case 1, random sample)

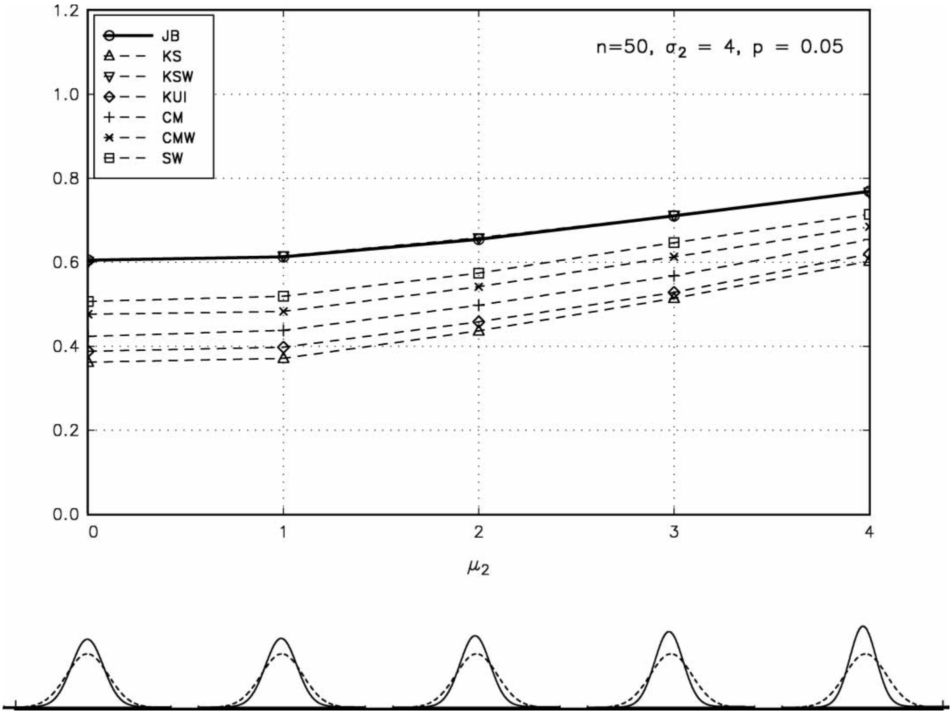

Figure 9. Power of the tests (asymmetric case 2, random sample)

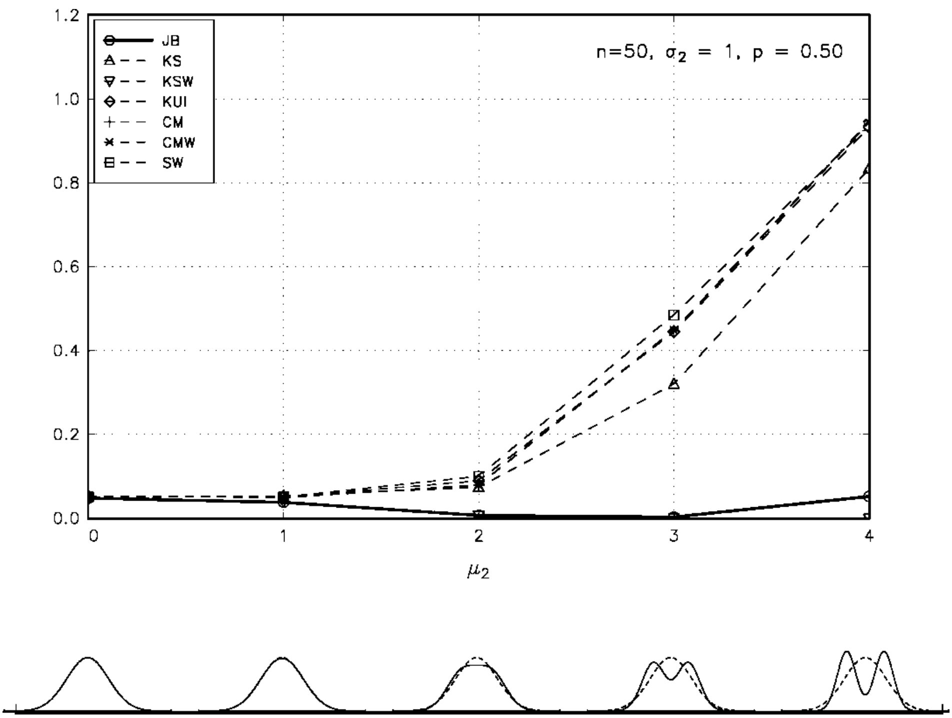

Figure 10. Power of the tests (symmetric case 4, random sample)

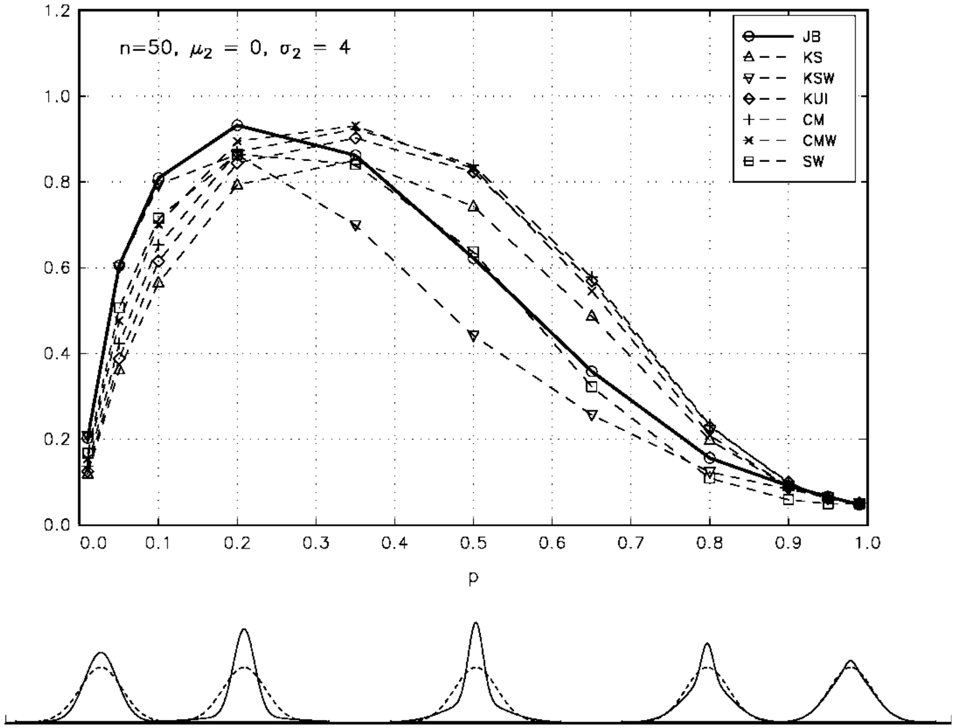

Figure 11. Power of the tests (symmetric case 5, random sample)

increasing values of σ_2 the kurtosis β_2 increases, for $p = 0.1$ much more than for $p = 0.5$ and 0.8. In the symmetric cases (Figures 5 to 7), we see that for small values of p the JB-test is the best one, the power of JB, however, decreases with increasing p where β_2 increases slower. For great values of p the Kuiper test and CMW-test are superior, the Shapiro–Wilk test is, in general, not a powerful test in comparison to the others. In the asymmetric case with $p = 0.5$, $\mu_2 = 3$ and $\sigma_2 \neq 1$ (Figure 8), where the skewness β_1 is smaller than 1, JB and KSW have lowest power, both tests are even biased for the symmetric case, where the kurtosis is equal to 2.04, smaller than that of the standard normal distribution ($\beta_2 = 3$).

Figures 9 and 10 present power curves as functions of μ_2, again, the kurtosis is much greater for $p = 0.05$ than for $p = 0.5$. In Figure 11 the power is plotted as a function of p with decreasing kurtosis.

Figure 9 shows that the JB-test is the best one for small $p = 0.05$ even in the asymmetric case but the power loss is dramatic for $p = 0.5$ (Figure 10), a distribution that is symmetric and bimodal with small values of β_2 (JB is biased). Here, the SW- and CMW-tests are the winner. For fixed $\mu_2 = 0$ and $\sigma_2 = 4$ (Figure 11) the JB-test behaves very well for values of p smaller, say, 0.25, but with increasing values of p, the power of JB decreases rapidly, the same is true for the other tests.

Further results on the power of tests on normality can be found in Chapter 7 of Thode (2002). At this point it should be noted that the kurtosis β_2 is not a suitable measure for tailweight, a discussion of this problem 'what does β_2 really measure?' can be found in Chapter 3.3 of Büning (1991) where measures of tailweight are given, e.g. the measure

$$T = \frac{x_{0.975} - x_{0.025}}{x_{0.875} - x_{0.125}}$$

Table 10. Values of kurtosis β_2 and tail measure T for symmetric CN

p	0	0.1	0.2	0.3	0.4	0.5	0.8
β_2	3.00	8.33	7.56	6.53	5.64	4.94	3.56
T	1.70	1.99	2.42	2.56	2.47	2.28	1.84

where x_p is the p-quantile of the distribution function F. Table 10 presents values of β_2 and T for symmetric contaminated normal distributions (CN) with $\mu_2 = 0$ and $\sigma_2 = 3$ varying values of p (see Figures 5 to 7).

We see that with increasing proportion p of contamination the kurtosis β_2 decreases whereas the tailweight T is, at first, increasing and then decreasing. That means, CN has, for example, longer tails for $p = 0.2$ than for $p = 0.1$, the kurtosis, however, is smaller for $p = 0.2$ than for $p = 0.1$.

Now, let us study the power of the tests in the case of regression variables. As already mentioned above we considered the cases of one, three and six independent regression variables assuming different distribution functions such as uniform, normal and exponential. It might be a surprising result that the pattern of the power curves of the tests in each sample situation (various numbers of variables and various distribution functions) is always nearly the same. Thus, we restrict the illustration to the case of three standard normally distributed regression variables, the power curves are given in Figure 12 for the symmetric case with $\mu_2 = 0$ and $p = 0.5$ and varying σ_2. Obviously, the JB-test behaves well, but the CMW-, KUI- and the KS-tests are superior with increasing σ_2.

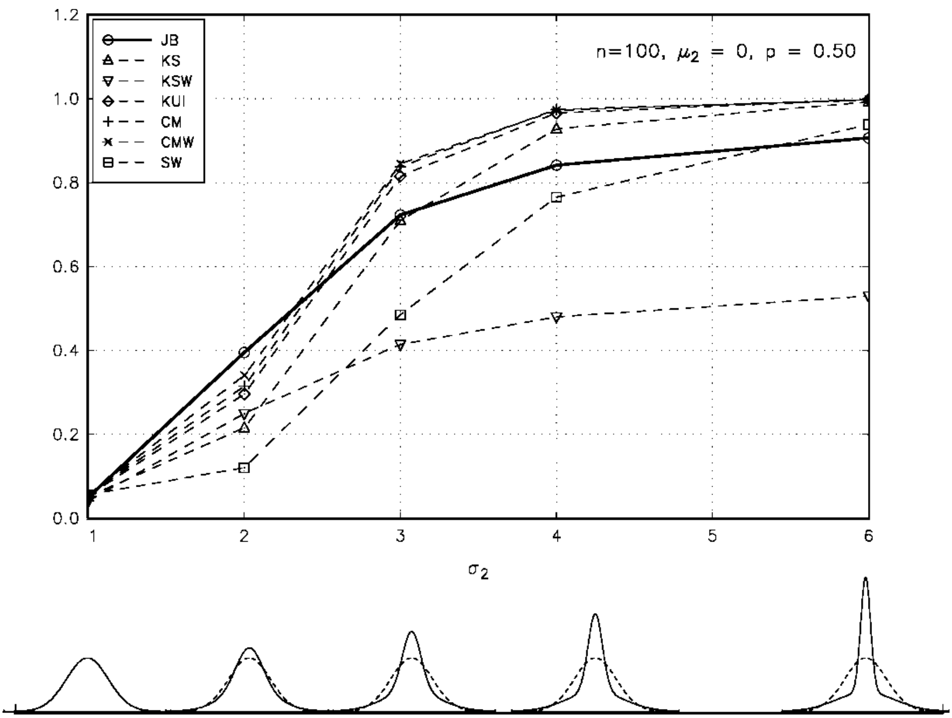

Figure 12. Power of the tests (symmetric case, three N(0,1)-regression variables)

For further studies of tests on normality in the linear regression model, see for example Huang & Bolch (1974) and White & MacDonald (1980).

We assumed independent regression variables. The question arises how robust are the tests above if the regression variables are correlated, an important problem for testing normality in regression analysis. A first study, assuming an autoregressive process of order one for the error terms with $\rho = -0.5$, shows that the influence on the critical values and therefore on the power of the tests is considerable. Further information about this topic is found, for example, in Bontemps & Meddahi (2005). There are a lot of references on other articles about testing normality in the context of time series.

Conclusions and Outlook

As an overall result of our power study we can state:

- The asymptotic JB-test is conservative for $\alpha = 0.05$, 0.10, the approximation by the Chi-square distribution does not work well.
- There are no remarkable differences between the critical values of the tests in model I with random samples and in model II with regression variables for $n \geq 50$.
- The modified version JBU of JB introduced by Urzúa (1996) means no improvement of power of the classical test.
- The non-parametric tests, KS and CM, as well as its modifications KSW and CMW, are very conservative in the case of estimating μ and σ if we use the critical values for the original data.
- The power values for each of the tests are nearly the same for random samples (model I) and for regression variables (model II).
- The JB-test behaves well (it is often even the best one) for symmetric distributions with medium up to long tails and for slightly skewed distributions with long tails.
- The power of the JB-test is poor for distributions with short tails, especially if the shape is bimodal, sometimes JB is even biased. In this case the modification of CM, CMW, or the Shapiro–Wilk test may be recommended.

Two further problems by testing normality in regression analysis are of theoretical and practical importance, the case of autocorrelated error terms as already mentioned above and the case of heteroscedastic error terms. The influence of autocorrelation on the critical values and the power of the tests seems to be strong as first studies show. Thus, the generalized least square estimator (GLS) is more appropriate for the regression parameters instead of the ordinary least square (OLS). For the case of heteroscedasticity of the error terms similar studies have to be done, too. For studies of serial correlation and heteroscedasticity of regression residuals, see for example Jarque & Bera (1980).

References

Anderson, T. & Darling, D. (1952) Asymptotic theory of goodness of fit criteria based on stochastic processes, *Annals of Mathematical Statistics*, 23, pp. 193–212.

Bontemps, C. &, Meddahi, N. (2005) Testing normality: a GGM approach, *Journal of Econometrics*, 124, pp. 149–186.

Bowman, K. & Shenton, L. (1975) Omnibus test contours for departures from normality based on b_1 and b_2, *Biometrika*, 62, pp. 243–250.

Büning, H. (1991) *Robuste und Adaptive Tests* (Berlin: Walter de Gruyter).

Büning, H. (2001) Kolmogorov–Smirnov and Cramér-von Mises type two-sample tests with various weight functions, *Communications in Statistics - Simulation and Computation*, 30, pp. 847–865.

Büning, H. & Trenkler, G. (1994) *Nichtparametrische Statistische Methoden* (Berlin: Walter de Gruyter).
Canner, P. (1975) A simulation study of one- and two-sample Kolmogorov–Smirnov statistics with a particular weight function, *Journal of the American Statistical Association*, 70, pp. 209–211.
D'Agostino, R. & Stephens, M. (1986) *Goodness-of-Fit-Techniques* (New York: Marcel Dekker).
Huang, C. & Bolch, B. (1974) On the testing of regression disturbances for normality, *Journal of the American Statistical Association*, 69, pp. 330–335.
Jarque, C. & Bera, A. (1980) Efficient tests for normality homoscedasticity and serial independence of regression residuals, *Econometric Letters*, 6, pp. 255–259.
Jarque, C. & Bera, A. (1987) A test for normality of observations and regression residuals, *International Statistical Review*, 55, pp. 163–172.
Kuiper, N. (1960) Tests concerning random points on a circle, *Nederl. Akad. Wetensch. Proc., Ser. A*, 63, pp. 38–47.
Shapiro, S. & Francia, R. (1972) An approximate analysis of variance test for normality, *Journal of the American Statistical Association*, 67, pp. 215–216.
Shapiro, S. & Wilk, M. (1965) An analysis of variance test for normality (complete samples), *Biometrika*, 52, pp. 591–611.
Shapiro, S., Wilk, M. & Chen, H. (1968) A comparative study of various tests for normality, *Journal of the American Statistical Association*, 63, pp. 1343–1372.
Stephens, M. (1974) Edf statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association*, 69, pp. 730–737.
Thode, H., Jr. (2002) *Testing Normality* (New York: Marcel Dekker).
Urzúa, C. (1996) On the correct use of omnibus tests for normality, *Economics Letters*, 53, pp. 247–251.
White, H. & MacDonald, G. (1980) Some large-sample tests for normality in the linear regression model, *Journal of the American Statistical Association*, 75, pp. 16–28.