International Statistical Review (1987), **55**, 2, pp. 163–172. Printed in Great Britain

A Test for Normality of Observations and Regression Residuals

Carlos M. Jarque[1] and Anil K. Bera[2]

[1] *Dirección General de Estadística, Secretaria de Programacion y Presupuesto, Mexico D.F.* [2] *Department of Economics, University of Illinois, Urbana-Champaign, Illinois 61820, USA*

Summary

Using the Lagrange multiplier procedure or score test on the Pearson family of distributions we obtain tests for normality of observations and regression disturbances. The tests suggested have optimum asymptotic power properties and good finite sample performance. Due to their simplicity they should prove to be useful tools in statistical analysis.

Key words: Lagrange multiplier test; Normality test; Regression model; Score test.

1 Introduction

Statisticians' interest in fitting curves to data goes a long way back. As noted by Ord (1972, p. 1), although towards the end of the nineteenth century 'not all were convinced of the need for curves other than the normal' (K. Pearson, 1905), 'by the turn of the century most informed opinion had accepted that populations might be non-normal'; some historical accounts are given by E. S. Pearson (1965). This naturally led to the development of tests for the *normality of observations*. Interest in this area is still very much alive, and recent contributions to the literature are the skewness, kurtosis and omnibus tests proposed by D'Agostino & Pearson (1973), Bowman & Shenton (1975) and Pearson, D'Agostino & Bowman (1977), the analysis of variance tests of Shapiro & Wilk (1965) and Shapiro & Francia (1972), and the coordinate-dependent and invariant procedures described by Cox & Small (1978).

There has also been considerable recent interest in testing the *normality of* (*unobserved*) *regression disturbances*. This is noted below, but first we introduce necessary notation. We consider the linear regression model, $y_i = x_i'\beta + u_i$, for $i = 1, \ldots, N$, where x_i' is a 1 by K vector of observations on K fixed regressors, β is a K by 1 vector of unknown parameters, and u_i is the ith unobservable disturbance assumed to have zero mean and to be homoscedastic, identically distributed and serially independent. An additional assumption frequently made in this model is that the probability density function of u_i, $f(u_i)$, is the normal probability density function.

The consequences of violation of this normality assumption have been studied by various authors. In estimation, for instance, the ordinary least-squares estimator $b = (X'X)^{-1}X'y$, which is known to be efficient under normality, may be very sensitive to long-tailed distributions; for example, see Hogg (1979). Regarding inferential procedures, Box & Watson (1962) consider the usual t and F-tests, and demonstrate that sensitivity to nonnormality is determined by the numerical values of the regressors. They also show that, to obtain the desired significance level, some adjustment in the degrees of freedom of these tests may be required. Similarly, Arnold (1980) studies the distribution of

$s^2 = (y - Xb)'(y - Xb)/N$ and shows that the significance level of the usual χ^2 test of the hypothesis $\sigma^2 = \sigma_0^2$ is not even asymptotically valid in the presence of nonnormality. Furthermore, it has been found that homoscedasticity and serial independence tests suggested for normal disturbances may result in incorrect conclusions under non-normality; for example, see Bera & Jarque (1982). In all, violation of the normality assumption may lead to the use of suboptimal estimators, invalid inferential statements and to inaccurate conclusions, highlighting the importance of testing the validity of the assumption.

In § 2, we present a procedure for the construction of efficient and computationally simple statistical specification tests. This is used in § 3 to obtain a test for the normality of observations, and in § 4 to obtain a test for the normality of (unobserved) regression disturbances. We then present in § 5 an extensive simulation study to compare the power of the tests suggested with that of other existing procedures.

2 The score test

The procedure we present for the construction of specification tests consists of the use of the score test on a 'general family of distributions'. The score test, also known as the Lagrange multiplier test, is fully described elsewhere, for example, see Cox & Hinkley (1974, Ch. 9), so here we only introduce notation and state required results.

Consider a random variable u with probability density function $f(u)$. For a given set of N independent observations on u, say $u_1, \ldots, u_N$, denote by $l(\theta) = l_1(\theta) + \ldots + l_N(\theta)$ the logarithm of the likelihood function, where $l_i(\theta) = \log f(u_i)$, $\theta = (\theta_1', \theta_2')'$ is the vector of parameters (of finite dimension), and θ_2 is of dimension r by 1. Assume we are interested in testing the hypothesis $H_0: \theta_2 = 0$.

Define $d_j = \sum \partial l_i(\theta)/\partial\theta_j$ and $\mathscr{I}_{jk} = E[\sum (\partial l_i(\theta)/\partial\theta_j)(\partial l_i(\theta)/\partial\theta_k)']$, where summation goes from $i = 1$ to N, for $j = 1, 2$ and $k = 1, 2$. Let $\hat{d}_j$ and $\hat{\mathscr{I}}_{jk}$ denote d_j and $\mathscr{I}_{jk}$ evaluated at the restricted (obtained by imposing the restriction $\theta_2 = 0$) maximum likelihood estimator of θ, say $\hat{\theta}$.

It may be shown that under general conditions, the statistic defined by

$$\text{LM} = \hat{d}_2'(\hat{\mathscr{I}}_{22} - \hat{\mathscr{I}}_{21}\hat{\mathscr{I}}_{11}^{-1}\hat{\mathscr{I}}_{12})^{-1}\hat{d}_2 \tag{1}$$

is, under $H_0: \theta_2 = 0$, asymptotically distributed as a χ^2 with r degrees of freedom, say $\chi^2_{(r)}$. A test of $H_0: \theta_2 = 0$, based on (1), will be referred to as a score or LM test. Two aspects of this test are worth nothing. *First,* that it is asymptotically equivalent to the likelihood ratio test, implying it has the same asymptotic power characteristics including maximum local asymptotic power (Cox & Hinkley, 1974). *Secondly,* that to compute it we only require estimation under the null hypothesis. In the inferential problems studied here, estimation under H_0 is easily carried out making the test computationally attractive, as compared to other asymptotically equivalent procedures (i.e., likelihood ratio and Wald tests). For these two reasons (*good power and computational ease*) we use the LM test, rather than others, in the inferential procedure.

It should be mentioned that the Lagrange multiplier method has been applied recently in many econometric problems; for example, see Byron (1970), Godfrey (1978) and Breusch & Pagan (1980). Here the procedure for the construction of specification tests also uses this principle, but has its distinct feature in the formulation of $l(\theta)$. Rather than assuming a 'particular' probability density function for u_i (or transformation of u_i), we assume that the true probability density function for u_i belongs to a 'general family' (for example, the Pearson family), of which the distribution under H_0 is a particular member.

We then apply the Lagrange multiplier principle to test H_0 within this 'general family' of distributions. Our subsequent discussion will help make this point clearer. The tests obtained are known to have optimal large sample power properties for members of the 'general family' specified. Yet, this does not imply they will not have good power properties for nonmember distributions. Indeed, as shown in § 5, the tests can perform with extremely good power even for distributions not belonging to the 'general family' from which the test was derived.

3 A test for normality of observations

In this section we make use of the Lagrange multiplier method to derive an additional test for the normality of observations which is simple to compute and asymptotically efficient.

Consider having a set of N independent observations on a random variable v, say $v_1, \ldots, v_N$, and assume we are interested in testing the normality of v. Denote the unknown population mean of v_i by $\mu = E[v_i]$ and, for convenience, write $v_i = \mu + u_i$. Assume the probability density function of $u_i, f(u_i)$, is a member of the *Pearson family*. This is not very restrictive, due to the wide range of distributions that are encompassed in it (e.g., particular members are the normal, beta, gamma, Student's t and F distributions). This means we can write (Kendall & Stuart, 1969, p. 148)

$$df(u_i)/du_i = (c_1 - u_i)f(u_i)/(c_0 - c_1 u_i + c_2 u_i^2) \quad (-\infty < u_i < \infty). \tag{2}$$

It follows that the logarithm of the likelihood function of our N observations $v_1, \ldots, v_N$ may be written as

$$l(\mu, c_0, c_1, c_2) = -N \log \left[\int_{-\infty}^{\infty} \exp \left[\int \frac{c_1 - u_i}{c_0 - c_1 u_i + c_2 u_i^2} du_i \right] du_i \right] + \sum_{i=1}^{N} \left[\int \frac{c_1 - u_i}{c_0 - c_1 u_i + c_2 u_i^2} du_i \right]. \tag{3}$$

Our interest here is to test the hypothesis of normality, which means, from our expression for $f(u_i)$, that we want to test $H_0: c_1 = c_2 = 0$. Let $\theta_1 = (\mu, c_0)'$, $\theta_2 = (c_1, c_2)'$ and $\theta = (\theta_1', \theta_2')'$. Using these, and the definitions of § 2, we can show that the LM test statistic is given by

$$\text{LM} = N[(\sqrt{b_1})^2/6 + (b_2 - 3)^2/24], \tag{4}$$

where $\sqrt{b_1} = \hat{\mu}_3/\hat{\mu}_2^{3/2}$, $b_2 = \hat{\mu}_4/\hat{\mu}_2^2$, $\hat{\mu}_j = \sum (v_i - \bar{v})^j/N$ and $\bar{v} = \sum v_i/N$, with summations going from $i = 1$ to N. Note that $\sqrt{b_1}$ and b_2 are, respectively, the skewness and kurtosis sample coefficients. From the results stated in § 2 we know that, under $H_0: c_1 = c_2 = \theta$, LM is asymptotically distributed as $\chi^2_{(2)}$, and that a test based on (4) is asymptotically locally most powerful. The hypothesis H_0 is rejected, for large samples, if the computed value of (4) is greater than the appropriate significance point of a $\chi^2_{(2)}$.

Several tests for normality of observations are available. For example, there are tests based on either of the quantities $\sqrt{b_1}$ or b_2. These have optimal properties, for large sample, if the departure from normality is due to either skewness or kurtosis (Geary, 1947). In addition, there are omnibus tests based on the joint use of $\sqrt{b_1}$ and b_2. One example is the R test suggested by Pearson et al. (1977); see also D'Agostino & Pearson (1973, p. 620).

It is interesting to note that (4) is the test given by Bowman & Shenton (1975). Bowman & Shenton (1975, p. 243) only stated the expression of the statistic, and noted it was asymptotically distributed as $\chi^2_{(2)}$ under normality; they did not study its large or finite sample properties. We have shown expression (4) is a score or LM test statistic. Therefore, we have uncovered a principle that proves its asymptotic efficiency. This finding encourages the study of its finite sample properties. For finite N, the distributions of $\sqrt{b_1}$ and b_2, under H_0, are still unknown. The problem has engaged statisticians for a number of years, and only approximations to the true distributions are available; for example, for $\sqrt{b_1}$ see D'Agostino & Tietjen (1973), and for b_2 see D'Agostino & Pearson (1973). This highlights, together with the fact that $\sqrt{b_1}$ and b_2 are not independent, for example, see Pearson et al. (1977, p. 233), the difficulty of analytically obtaining the finite sample distribution of (4) under H_0.

An alternative is to resort to computer simulation. We see that LM is invariant to the scale parameter, i.e., that the value of LM is the same if computed with v_i/σ rather than v_i (for all finite $\sigma > 0$). Therefore, we may assume $V[v_i] = 1$, and generate n sets of N pseudo-random variates from a $N(0, 1)$. Then, for each of these n sets, LM would be computed, giving n values of LM under H_0. By choosing n large enough, we may obtain as good an approximation as desired to the distribution of LM and, so, determine the critical point of the test for a given significance level α, or the probability of a type I error for the computed value of LM from a particular set of observations. Computer simulation is used in § 5.1. There, we present a study comparing the finite sample power of LM with that of other existing tests for normality, and a table of significance points for $\alpha = 0{\cdot}10$ and $0{\cdot}05$.

To finalize this section, we note that the procedure utilized here may be applied in a similar way to other families of distributions. We have used the *Gram–Charlier* (*type A*) *family,* for example, see Kendall & Stuart (1969, p. 156), and derived the LM normality test, obtaining the same expression as for the *Pearson family,* i.e. equation (4). Our approach may also be used to test the hypothesis that $f(u)$ is any particular member of, say, the *Pearson family.* This may be done by forming H_0 with the appropriate values of c_0, c_1 and c_2 that define the desired distribution, e.g., to test if $f(u)$ is a gamma distribution we would test $H_0: c_1 = 0$ (Kendall & Stuart, 1969, p. 152). In some cases this may involve testing nonlinear inequalities in c_0, c_1 and c_2. For example, to test if $f(u)$ is a Pearson type IV we would test $H_0: c_1^2 - 4c_0c_2 < 0$. This requires the development of the score or Lagrange multiplier procedure to test nonlinear inequalities, and is an area for further research.

4 A test for normality of disturbances

Now we consider the regression model given in § 1. We note that the regression disturbances $u_1, \ldots, u_N$ are assumed to be independent and identically distributed with population mean equal to zero. In addition, we now assume that the probability density function of u_i, $f(u_i)$, is a member of the *Pearson family*; the same result is obtained if we use the *Gram–Charlier* (*type A*) *family.* This means we can define $f(u_i)$ as in (2) and the log-likelihood of our N observations $y_1, \ldots, y_N$ as in (3) where now the parameters, i.e. the arguments in $l(.)$, are β, c_0, c_1 and c_2, and $u_i = y_i - x_i'\beta$.

We define $\theta_1 = (\beta', c_0)'$ and $\theta_2 = (c_1, c_2)'$. To test the normality of the disturbances is equivalent to testing $H_0: \theta_2 = 0$. It can be shown that, in this case, the test statistic becomes

$$\mathrm{LM}_{\mathcal{N}} = N[\hat{\mu}_3^2/(6\hat{\mu}_2^3) + ((\hat{\mu}_4/\hat{\mu}_2^2) - 3)^2/24] + N[3\hat{\mu}_1^2/(2\hat{\mu}_2) - \hat{\mu}_3\hat{\mu}_1/\hat{\mu}_2^2], \tag{5}$$

where now $\hat{\mu}_j = \sum \hat{u}_i^j/N$, with the sum over $i = 1, \ldots, N$, and the $\hat{u}_i$ are the ordinary least-squares residuals, that is, $\hat{u}_i = y_i - x_i'b$. We have written the resulting test statistic with a suffix $\mathcal{N}$ to indicate this refers to a *disturbance normality* test. Using the results of

§ 2, we know $\text{LM}_{\mathcal{N}}$ is, under H_0, asymptotically distributed as $\chi^2_{(2)}$, and that it is asymptotically efficient. Obtaining the finite sample distribution of $\text{LM}_{\mathcal{N}}$ by analytical procedures appears to be intractable. For a given matrix X, we may resort to computer simulation, generating u_i from a $N(0, 1)$; for example, see § 3 and note $\text{LM}_{\mathcal{N}}$ is invariant to the scale parameter σ^2. In § 5.2 we use computer simulation to study the finite sample power of $\text{LM}_{\mathcal{N}}$.

To finalize, we recall that, in linear models with a constant term, ordinary least-squares residuals satisfy the condition $\hat{u}_1 + \ldots + \hat{u}_N = 0$. In these cases we have $\hat{\mu}_1 = 0$ and, therefore, (5) would reduce to

$$\text{LM}_{\mathcal{N}} = N[(\sqrt{\hat{b}_1})^2/6 + (\hat{b}_2 - 3)^2/24], \tag{6}$$

where $\hat{b}_1 = \hat{\mu}_3^2/\hat{\mu}_2^3$ and $\hat{b}_2 = \hat{\mu}_4/\hat{\mu}_2^2$. So we see that, regardless of the existence or not of a constant term in the regression, the statistic is extremely simple to compute requiring only the first four sample moments of the ordinary least-squares residuals.

5 Power of normality tests

In this section we present results of a *Monte Carlo study* done to compare the power of various tests for normality of observations and regression disturbances. We carried out simulations for small and moderate sample sizes; more specifically, we used $N = 20$, 35, 50, 100, 200 and 300. We consider four distributions from the Pearson family: the normal, gamma (2, 1), beta (3, 2) and Student's t with 5 degrees of freedom; and one nonmember: the lognormal. These distributions were chosen because they cover a wide range of values of third and fourth standardized moments (Shapiro, Wilk & Chen, 1968, p. 1346). To generate pseudo-random variates u_i, from these and other distributions considered throughout the study, we used the subroutines described by Naylor et al. (1966) on a UNIVAC 1100/42. Each of the five variates mentioned above was standardized so as to have zero mean.

5.1 *Testing for normality of observations*

We first note that, since $\mu = 0$, we have $v_i = u_i$; see § 3 for notation. The tests we consider for the normality of the observations u_i have the following forms:

(i) skewness measure test: reject normality, that is H_0, if $\sqrt{b_1}$ is outside the interval $(\sqrt{b_{1L}}, \sqrt{b_{1U}})$;

(ii) kurtosis measure test: reject H_0 if b_2 is outside (b_{2L}, b_{2U});

(iii) D'Agostino (1971) D^* test: reject H_0 if

$$D^* = [\Sigma\,(i/N^2 - (N+1)/(2N^2))e_i^0/\mu_2^{\frac{1}{2}} - (2\sqrt{\pi})^{-1}]N^{\frac{1}{2}}/0{\cdot}02998598$$

is outside (D_L^*, D_U^*), where e_i^0 is the ith order statistic of $u_1, \ldots, u_N$;

(iv) Pearson et al. (1977) R test: reject H_0 if either $\sqrt{b_1}$ is outside (R_{1L}, R_{1U}) or b_2 is outside (R_{2L}, R_{2U});

(v) Shapiro & Wilk (1965) W test: reject H_0 if $W = (\Sigma\, a_{iN}e_i^0)^2/(N\mu_2)$ is less than W_L, where the a_{iN} are coefficients tabulated by Pearson & Hartley (1972, p. 218) or obtainable following Royston (1982 a, b);

(vi) Shapiro & Francia (1972) W' test: reject H_0 if $W' = (\Sigma\, a'_{iN}e_i^0)^2/(N\mu_2)$ is less than W'_L, where the a'_{iN} are coefficients that may be computed using the tables of Harter (1961); and

(vii) LM test: reject H_0 if $\text{LM} > \text{LM}_U$.

We did not include *distance tests* (such as the Kolmogorov–Smirnov test, Cramér-von Mises test, weighted Cramér–von Mises test and the Durbin test) because it has previously been reported that, for a wide range of alternative distributions, the W test, considered here, was superior to these (Shapiro et al., 1968).

The values $\sqrt{b_{1L}}$, $\sqrt{b_{1U}}$, b_{2L}, b_{2U}, D_L^*, D_U^*, R_{1L}, R_{1U}, R_{2L}, R_{2U}, W_L, W_L' and LM_U are appropriate significance points. We considered a 10% significance level; i.e., we set $\alpha = 0{\cdot}10$. For $N = 20$, 35, 50 and 100 and tests $\sqrt{b_1}$, b_2, D^* and R, the points are as given by White & MacDonald (1980, p. 20). For $n = 200$ and 300, significance points for $\sqrt{b_1}$, b_2 and D^* were obtained respectively from Pearson & Hartley (1962, p. 183), Pearson & Hartley (1962, p. 184) and D'Agostino (1971, p. 343); and for the R test we extrapolated the points for $N \leqslant 100$. For W, W' and LM we computed the significance points by simulation using 250 replications so that the empirical α, say $\hat{\alpha}$, was equal to 0·10. For example, for a given N, we set $W_L = W(25)$, where $W(25)$ was the 25th largest of the values of W in the 250 replications under normal observations. Similarly for W'. For LM we set $\text{LM}_U = \text{LM}(225)$. Initially we used, for W, the points from Shapiro & Wilk (1965, p. 605); for W' from Weisberg (1974, p. 645) and Shapiro & Francia (1972, p. 216); and for LM from the values of Table 2. With $\hat{\alpha} = 0{\cdot}10$, easier power comparisons among the one-sided tests W, W' and LM can be made. Note that $\sqrt{b_1}$, b_2 and D^* are two-sided tests and that R is a four-sided test, and hence, for these, it is troublesome to adjust the significance points so that $\hat{\alpha} = 0{\cdot}10$.

Every experiment in this simulation study consists of generating N pseudo-random variates from a given distribution; computing the values of $\sqrt{b_1}$, b_2, D^*, W, W' and LM; and seeing whether H_0 is rejected by each individual test. We carried out 250 replications and estimated the power of each test by dividing by 250 the number of times H_0 was rejected. In Table 1 we present the power calculation obtained. For economy of presentation, we only report full numerical results for sample sizes $N = 20$ and $N = 50$, which are sufficient for the purpose of arriving at conclusions. It should be mentioned that, given the number of replications, the power calculations reported in Table 1 (and also in Table 3) are accurate to only two decimal points (with a 90% confidence coefficient). This is not regarded as unsatisfactory for the objective of our analysis; see David (1970, Ch. 2) for the derivation of confidence intervals for quantiles, which we used to obtain intervals for the 10% critical points of the tests.

If we have large samples, and we are considering members of the Pearson family, the theoretical results of § 3 justify the use of the LM test. For finite sample performance we resort to Table 1. Here we see that the preferred tests would probably be W and LM, followed by W'. It is also interesting to note that, for $N = 100$, 200 and 300, LM had the highest power for all distributions, but differences with the W and W' tests were small; for example for $N = 200$, power for LM was 0·964, 0·856, 1·0 and 1·0 which compares with 0·944, 0·848, 1·0 and 1·0 for W', respectively, for the beta, Student's t, gamma and

Table 1
Normality of observations; *estimated power with* 250 *replications*; $\alpha = 0{\cdot}10$

N		$\sqrt{b_1}$	b_2	D^*	R	W	W'	LM
20	Beta	0·072	0·128	0·124	0·120	0·208	0·132	0·116
	Student's t	0·272	0·212	0·240	0·252	0·280	0·300	0·340
	Gamma	0·796	0·476	0·604	0·772	0·920	0·884	0·872
	Lognormal	0·996	0·916	0·988	0·996	0·996	0·996	0·996
50	Beta	0·192	0·232	0·204	0·292	0·480	0·360	0·412
	Student's t	0·372	0·404	0·404	0·420	0·332	0·496	0·508
	Gamma	1·000	0·768	0·920	1·000	1·000	1·000	1·000
	Lognormal	1·000	1·000	1·000	1·000	1·000	1·000	1·000

Table 2

Normality of observations; significance points for LM *normality test*; 10 000 *replications*

N	$\alpha = 0{\cdot}10$	$\alpha = 0{\cdot}05$	N	$\alpha = 0{\cdot}10$	$\alpha = 0{\cdot}05$
20	2·13	3·26	200	3·48	4·43
30	2·49	3·71	250	3·54	4·51
40	2·70	3·99	300	3·68	4·60
50	2·90	4·26	400	3·76	4·74
75	3·09	4·27	500	3·91	4·82
100	3·14	4·29	800	4·32	5·46
125	3·31	4·34	∞	4·61	5·99
150	3·43	4·39			

lognormal densities. Additionally we found that LM may have good relative power even when the distribution is not a member of the Pearson family; for example, see power for lognormal in Table 1. Overall, LM *is preferred,* followed by W and W', which in turn dominate the other four tests.

Apart from power considerations, LM has an *advantage* over W (and W') in that, for its computation, one requires neither ordered observations (which may be expensive to obtain for large N) nor expectations and variances and covariances of standard normal order statistics. The simulation results presented, together with its proven asymptotic properties suggest the LM test may be the preferred test in many situations. Therefore, it appeared worthwhile to carry out extensive simulations to obtain, under normality, finite sample significance points for LM. Using expression (4) we carried out 10 000 replications and present, in Table 2, significance points for $\alpha = 0{\cdot}10$ and 0·05 for a range of sample sizes. This table should be useful in applications of the test.

5.2 *Testing for normality of regression disturbances*

We now study the power of tests for normality of (unobserved) regression disturbances. The tests we consider are the same as those described in § 5.1, but we computed them with estimated regression residuals rather than the true disturbances u_i. We denote these by $\sqrt{\hat{b}_1}$, $\hat{b}_2$, $\hat{D}^*$, $\hat{R}$, $\hat{W}$, $\hat{W}'$ and $\text{LM}_{\mathcal{N}}$. The first six are the modified large-sample tests discussed by White & MacDonald (1980). The seventh test is the one suggested in § 4. The modified Shapiro–Wilk test, $\hat{W}$, has been reported to be superior to modified distance tests, so these were excluded (Huang & Bolch, 1974, p. 334).

We consider a linear model with a constant term and three additional regressors, i.e. with $K = 4$, and utilize the ordinary least-squares residuals $\hat{u}_i$ to compute the modified tests. Huang & Bolch (1974) and Ramsey (1974, p. 36) have found that the power of modified normality tests, computed using ordinary least-squares residuals, is higher than when using Theil's (1971, p. 202) best linear unbiased scalar covariance matrix residuals; see also Pierce & Gray (1982) for other possible scalings of the residuals. To obtain $\hat{u}_i$ we use the same u_i's as those generated in § 5.1. For comparison purposes, our regressors $X_1, \ldots, X_K$ are defined as by White & MacDonald (1980, p. 20); that is, we set $X_{i1} = 1$ $(i = 1, \ldots, N)$ and generate X_2, X_3 and X_4 from a uniform distribution. Note that the specific values of the means and variances of these regressors have no effect on the simulation results. This invariance property follows from the fact that, for a linear model with regressor matrix $X = (x_1, \ldots, x_N)'$ the ordinary least-squares residuals are the same as those of a linear model with regressor matrix XR, where R is any K by K nonsingular matrix of constants (Weisberg, 1980, p. 29). For $N = 20$ we use the first 20 of the 300 (generated) observations x_i. Similarly for $n = 35$, 50, 100 and 200.

For this part of the study we utilize the same significance points as those of § 5.1, except

Table 3

Normality of disturbances; estimated power with 250 *replications*; $\alpha = 0{\cdot}10$, $K = 4$; *regressors*, $X_1 = 1$; X_2, X_3, $X_4 \sim$ *uniform*; *N varies*

N		$\sqrt{\hat{b}_1}$	$\hat{b}_2$	$\hat{D}^*$	$\hat{R}$	$\hat{W}$	$\hat{W}'$	$\text{LM}_{\mathcal{N}}$
20	Beta	0·068	0·108	0·096	0·100	0·124	0·072	0·084
	Student's t	0·224	0·192	0·188	0·204	0·168	0·192	0·256
	Gamma	0·640	0·356	0·416	0·527	0·644	0·600	0·644
	Lognormal	0·920	0·844	0·904	0·912	0·924	0·932	0·944
50	Beta	0·160	0·180	0·148	0·212	0·344	0·172	0·188
	Student's t	0·360	0·388	0·400	0·412	0·300	0·456	0·464
	Gamma	0·984	0·724	0·856	0·976	0·988	0·988	0·988
	Lognormal	1·000	1·000	1·000	1·000	1·000	1·000	1·000

for $\hat{W}$, $\hat{W}'$ and $\text{LM}_{\mathcal{N}}$, for which we use the points corresponding to $\hat{\alpha} = 0{\cdot}10$; for example, as significance point of $\hat{W}$ we use $\hat{W}(25)$, where $\hat{W}(25)$ is the 25th largest of the values of $\hat{W}$ in the 250 replications under normal disturbances. The estimated power of each test for $N = 20$ and 50 is given in Table 3.

For $N = 20$, 35 and 50 we found that probably the best tests were $\text{LM}_{\mathcal{N}}$ and $\hat{W}$, followed by $\hat{W}'$. For $N = 100$, 200 and 300, $\text{LM}_{\mathcal{N}}$ had highest power for all distributions and $\hat{W}$ and $\hat{W}'$ performed quite well also; for example, for $N = 200$ power for $\text{LM}_{\mathcal{N}}$ was 0·928, 0·828, 1·0 and 1·0 which compares with 0·924, 0·820, 1·0 and 1·0 for $\hat{W}'$, respectively, for the beta, Student's t, gamma and lognormal densities. Our results agree with those of White & MacDonald (1980) in that, almost all the cases, the modified tests gave, correspondingly, lower powers than those using the original disturbances with differences diminishing as n increases; compare Tables 1 and 3. We also found that, for a given N and a given distribution, the ranking of the tests in both tables was approximately the same. To obtain a measure of closeness between the true and modified statistics we computed their correlation. The numerical results are not reported, but our findings agreed with those of White & MacDonald (1980, p. 22): $\sqrt{\hat{b}_1}$ appears to be closer to $\sqrt{b_1}$; and $\hat{b}_2$ appears to be closer to b_2, than the other modified statistics. In our study, these would be followed by $(D^*, \hat{D}^*)$ and then by $(\text{LM}, \text{LM}_{\mathcal{N}})$. We would then have $(W', \hat{W}')$ and, lastly, $(W, \hat{W})$.

A further comment is required. It is clear that the ordinary least-squares residuals $\hat{u} = (I - Q_X)u$ are a linear transformation (defined by Q_X) of the unobserved disturbances u, where $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_N)'$, $u = (u_1, \ldots, u_N)'$ and Q_X is an N by N matrix defined by $Q_X = X(X'X)^{-1}X'$. As noted by White & MacDonald (1980) and Weisberg (1980), simulation results studying the relative power of tests for the normality of u, computed using $\hat{u}$, depend on the particular form of Q_X. If one is to carry out a Monte Carlo study then, to have a less restrictive result, one should consider various forms of Q_X. Different forms may arise due to changes in N; due to variations in the way the regressors $X_1, \ldots, X_K$ are generated; and/or due to changes in the number of regressors K.

So far we have studied the power of the tests for different values of N, using $K = 4$ and generating the regressors as White & MacDonald (1980). In addition, we have repeated our experiments but generating the regressors in a different way. We set $X_{i1} = 1$ $(i = 1, \ldots, N)$ and generated X_2 from a normal, X_3 from a uniform and X_4 from a $\chi^2_{(10)}$. These regressor-distributions are of interest since they cover a convenient range of alternatives. The numerical results are not presented in detail, but our findings do not vary substantially from those stated for the White & MacDonald regressor set. For $N \leqslant 50$, $\text{LM}_{\mathcal{N}}$ is a preferred test (together with $\hat{W}$ and $\hat{W}'$), and is preferable to all tests for $N \geqslant 100$. The conclusions from the analysis of the correlations between the true and modified statistics are also the same.

As a final exercise, we carried out our experiment fixing $N=20$ and using the three regressor data sets reported by Weisberg (1980, p. 29). Following Weisberg, we varied K, for each data set, using $K=4$, 6, 8 and 10. Weisberg found that the power of the $\hat{W}'$ test may vary as K and/or the regressors are changed. We find this to be the case for all the tests considered. For example, for data set 1 we obtained that for the lognormal the power of $\sqrt{\hat{b}_1}$, say $P(\sqrt{\hat{b}_1})$, was equal to 0·636 with $K=10$ and to 0·956 with $K=4$; that is, $0{\cdot}636 \leqslant P(\sqrt{\hat{b}_1}) \leqslant 0{\cdot}956$. Similarly we obtained $0{\cdot}572 \leqslant P(\hat{b}_2) \leqslant 0{\cdot}812$, $0{\cdot}608 \leqslant P(\hat{D}^*) \leqslant 0{\cdot}888$, $0{\cdot}632 \leqslant P(\hat{R}) \leqslant 0{\cdot}940$, $0{\cdot}592 \leqslant P(\hat{W}) \leqslant 0{\cdot}928$, $0{\cdot}636 \leqslant P(\hat{W}') \leqslant 0{\cdot}944$ and $0{\cdot}676 \leqslant P(\mathrm{LM}_{\mathcal{N}}) \leqslant 0{\cdot}952$. We also found that for data sets 1 and 2, the empirical significance level $\hat{\alpha}$ is close to 0·10 for all statistics and all K. For data set 3, however, $\hat{\alpha}$ increased considerably as K increased; for example, for $\sqrt{\hat{b}_1}$, $\hat{\alpha}=0{\cdot}088$, 0·104, 0·200 and 0·216 respectively for $K=4$, 6, 8 and 10. This shows that the power and the level of a test may depend on the specific form of Q_X. Nevertheless, when comparing the relative power it was interesting to note that, for all K and all three regressor data sets, $\mathrm{LM}_{\mathcal{N}}$, $\hat{W}$ and $\hat{W}'$ were the preferred tests, as found in our earlier 2 sets of experiments with $N=20$. Regarding the correlations between the true and modified statistics, we observed that (for each data set) as K increased, all correlations decreased; for example, for data set 1, the correlation between D^* and $\hat{D}^*$, for the normal, was equal to 0·640 with $K=4$ and to 0·329 with $K=10$. However, the ranking among the tests remained the same.

In all, the results show that for all the forms of matrices Q_X studied, $\mathrm{LM}_{\mathcal{N}}$ performed with good relative power. This was true for both small N, for example, $N=20$, and large N, for example, $N=300$, encouraging the use of $\mathrm{LM}_{\mathcal{N}}$ when testing for the normality of regression disturbances. The statistic $\mathrm{LM}_{\mathcal{N}}$ is simple to compute and, in any regression problem, we may easily obtain an approximation to its finite sample distribution, under H_0, by computer simulation. This should not represent a serious problem, particularly with the fast speed and increased availability of modern computers. (A FORTRAN subroutine for the computation of the finite sample distribution of $\mathrm{LM}_{\mathcal{N}}$ is available from the authors upon request.) In turn, if one has large samples then comparison can be made with the critical point from a $\chi^2_{(2)}$.

Acknowledgements

The authors express their gratitude to the editor, associate editor and to the referee for suggestions which benefitted the quality of the paper.

References

Arnold, S.F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. *J. Am. Statist. Assoc.* **75**, 890–894.

Bera, A.K. & Jarque, C. M. (1982). Model specification tests: A simultaneous approach. *J. Econometrics* **20**, 59–82.

Bowman, K.O. & Shenton, L.R. (1975). Omnibus contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika* **62**, 243–250.

Box, G.E.P. & Watson, G.S. (1962). Robustness to non-normality of regression tests. *Biometrika* **49**, 93–106.

Breusch, T.S. & Pagan, A.R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Rev. Econ. Studies* **97**, 239–253.

Byron, R.P. (1970). The restricted Aitken estimation of sets of demand relations. *Econometrica* **38**, 816–830.

Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. New York: Wiley.

Cox, D.R. & Small, J.H. (1978). Testing multivariate normality. *Biometrika* **65**, 263–272.

D'Agostino, R.B. (1971). An omnibus test for normality for moderate and large size samples. *Biometrika* **58**, 341–348.

D'Agostino R.B. & Pearson, E.S. (1973). Tests for departure from normality: Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika* **60**, 613–622. Correction (1974), **61**, 647.

D'Agostino R.B. & Tietjen, G.L. (1973). Approaches to the null distribution of $\sqrt{b_1}$. *Biometrika* **60**, 169–173.

David, H.A. (1970). *Order Statistics*. New York: Wiley.
Geary, R.C. (1947). Testing for normality. *Biometrika* **34,** 209–242.
Godfrey, L.G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica* **46,** 1303–1310.
Harter, H.L. (1961). Expected values of normal order statistics. *Biometrika* **48,** 151–165.
Hogg, R.V. (1979). An introduction to robust estimation. In *Robustness in Statistics,* Ed. R.L. Launer and G.N. Wilkinson, pp. 1–17. New York: Academic Press.
Huang, C.J. & Bolch, B.W. (1974). On the testing of regression disturbances for normality. *J. Am. Statist. Assoc.* **69,** 330–335.
Kendall, M.G. & Stuart, A. (1969). *The Advanced Theory of Statistics,* **1,** London: Griffin.
Naylor, T.H., Balintfy, J.L., Burdick, D.S. & Chu, K. (1966). *Computer Simulation Techniques*. New York: Wiley and Sons.
Ord, J.K. (1972). *Families of Frequency Distributions,* Griffin's Statistical Monographs and Courses 30. London: Griffin.
Pierce, D.A. & Gray, R.J. (1982). Testing normality of errors in regression models. *Biometrika* **69,** 233–236.
Pearson, E.S. (1965). Studies in the history of Probability and Statistics, XIV: Some incidents in the early history of biometry and statistics, 1890–94. *Biometrika* **52,** 3–18.
Pearson, E.S. D'Agostino, R.B. & Bowman, K.O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika* **64,** 231–246.
Pearson, E.S. & Hartley, H.O. (1962). *Biometrika Tables for Statisticians,* **1**. Cambridge University Press.
Pearson, E.S. & Hartley, H.O. (1972). *Biometrika Tables for Statisticians,* **2**. Cambridge University Press.
Pearson, K. (1905). On the general theory of skew correlation and non-linear regression. *Biometrika* **4,** 172–212.
Ramsey, J.B. (1974). Classical model selection through specification error tests. In *Frontiers in Econometrics,* Ed. P. Zarembka, pp. 13–47. New York: Academic Press.
Royston, J.P. (1982a). An extension of Shapiro and Wilk's *W* test for normality to large samples. *Appl. Statist.* **2,** 115–124.
Royston, J.P. (1982b). Expected normal order statistics (exact and approximate). *Appl. Statist.* **3,** 161–165.
Shapiro, S.S. & Francia, R.S. (1972). An approximate analysis of variance test for normality. *J. Am. Statist. Assoc.* **67,** 215–216.
Shapiro, S.S. & Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52,** 591–611.
Shapiro, S.S., Wilk, M.B. & Chen, H.J. (1968). A comparative study of various tests for normality. *J. Am. Statist. Assoc.* **63,** 1343–1372.
Theil, H. (1971). *Principles of Econometrics*. Amsterdam: Wiley.
Weisberg, S. (1974). An empirical comparison of the percentage points of *W* and *W'*. *Biometrika* **61,** 644–646.
Weisberg, S. (1980). Comment on paper by H. White and G.M. MacDonald. *J. Am. Statist. Assoc.* **75,** 28–31.
White, H. & MacDonald, G.M. (1980). Some large sample tests for non-normality in the linear regression model. *J. Am. Statist. Assoc.* **75,** 16–28.

Résumé

En utilisant la procédure du multiplicateur de Lagrange, ou le score test, sur les distributions du genre Pearson, on obtient des tests de normalité pour les observations et les résidus de régression. Les tests suggérés ont des proprietés optimales asymptotiques et des bonnes performances pour des échantillons finis. A cause de leur simplicité ces tests doivent s'avérer comme des instruments utiles dans l'analyse statistique.

[*Received June* 1985, *revised May* 1986]