Econometrics Journal (1998), volume **1**, pp. 154–173.

Simulation-based finite sample normality tests in linear regressions

JEAN-MARIE DUFOUR[1], ABDELJELIL FARHAT[1], LUCIEN GARDIOL[1], LYNDA KHALAF[2]

[1]*C.R.D.E, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7*
E-mail: dufour@plgcn.umontreal.ca; farhat@dms.montreal.sc
[2]*GREEN, Université Laval, Pavillon J.-A.-DeSève, Ste. Foy, Québec, Canada G1K 7P4*
E-mail: lynda.khalaf@ecn.ulaval.ca

Received: June 1997, in final form April 1998

Summary In the literature on tests of normality, much concern has been expressed over the problems associated with residual-based procedures. Indeed, the specialized tables of critical points which are needed to perform the tests have been derived for the location-scale model; hence, reliance on available significance points in the context of regression models may cause size distortions. We propose a general solution to the problem of controlling the size of normality tests for the disturbances of standard linear regressions, which is based on using the technique of Monte Carlo tests. We study procedures based on 11 well-known test statistics: the Kolmogorov–Smirnov, Anderson–Darling, Cramér–von Mises, Shapiro–Wilk, Jarque–Bera and D'Agostino criteria. Evidence from a simulation study is reported showing that the usual critical values lead to severe size problems (over-rejections or under-rejections). In contrast, we show that Monte Carlo tests achieve perfect size control for any design matrix and have good power.

Keywords: *Normality test; Linear regression; Exact test; Monte Carlo test; Bootstrap; Kolmogorov–Smirnov; Anderson–Darling; Cramér–von Mises; Shapiro–Wilk; Jarque–Bera; D'Agostino.*

1. INTRODUCTION

The problem of testing normality is fundamental in both theoretical and empirical research. Indeed, the validity of parametric statistical inference procedures in finite samples (in the sense that their size is controlled) depends crucially on the underlying distributional assumptions. Consequently, there has been extensive focus on whether hypothesized distributions are compatible with the data. Tests of normality are particularly prevalent because the assumption of normality is quite often made in statistical analysis, e.g. in econometric studies. In this respect, the reviews by Mardia (1980), D'Agostino and Stephens (1986, Ch. 9) and Baringhaus *et al*. (1989) report nearly 40 different normality tests. For illustrative examples, see Fama (1976), Lee (1982), Bera *et al*. (1984), Harris (1986), Afflecks-Graves and McDonald (1989), Hall (1990), Richardson and Smith (1993), among others.

This paper will emphasize procedures applicable in the linear regression framework. We specifically address the problem of obtaining valid tests of disturbance normality based on several statistics already proposed in the literature. Eleven of the leading statistics are considered: (i) Kolmogorov–Smirnov, (ii) Anderson–Darling, (iii) Cramér–von Mises, (iv) Shapiro–Wilk, (v) Shapiro–Francia, (vi) Weisberg–Bingham, (vii) D'Agostino, (viii) Filliben, and (ix) Jarque–Bera; for a survey and references, see D'Agostino and Stephens (1986). These well-known tests have non-standard null distributions. Thus, tables of approximate critical points are provided for reference in practical applications. As most tables are derived from Monte Carlo calculations according to the location-scale model with independent and identically distributed (*i.i.d.*) observations, the problem of adequate use in regression contexts has long been recognized.

It was shown by Pierce and Kopecky (1979) that standard tests of normality (which account for an unknown mean and variance) are asymptotically valid when computed from regression residuals. These authors essentially studied the convergence of the empirical process of residuals. In location-scale contexts, the asymptotics of empirical processes and associated tests are well understood; see, for example, Durbin (1973a, b), Stephens (1976) and Pollard (1984). With respect to the regression model, Pierce and Kopecky have proved that the limiting process is the same for the least-squares residuals case as for *i.i.d.* observations. Consequently, statistics based on the sample process of residuals have the same asymptotic null distribution as in the location-scale model. Related findings were obtained independently by Loynes (1980) and Mukantseva (1977); see also Meester and Lockhart (1988) for a discussion of the case of designs with many blocks. These conclusions are based on finite dimensional asymptotics. In contrast, Mammen (1996) reconsidered the limiting behavior of tests of fit and the underlying processes allowing the dimension of the model to increase with the sample size. This author showed that in such a setting, residuals-based goodness-of-fit (GOF) procedures may break down in the following sense: even if the null hypothesis is true, standard tests tend to reject with high probability. Further recent results on empirical processes and associated tests in more general econometric models are available in Andrews (1988a, b, 1994).

The finite sample performance of regression-based normality tests has also received attention in the literature. From Monte Carlo experiments, Huang and Bolch (1974) and White and MacDonald (1980) concluded that computation of normality tests from residuals does not invalidate them. Yet Pierce and Gray (1982) and Weisberg (1980) have pointed out difficulties with the representativeness of this result and recommend the use of considerable caution in practical applications. These authors emphasize that reported Monte Carlo results depend crucially on specific experimental settings. The number of regressors, the sample size and the design matrix can all affect the validity of residual-based tests, in the sense that size distortions are quite likely (see the comments on the multiple regression case in D'Agostino and Stephens (1986, Section 9.6)). Similar concerns about size control are expressed by Poirier *et al*. (1986), Jarque and Bera (1987), Pfaffenberger and Dielman (1991) and Anderson (1994). Indeed, to obtain a valid power study, Pfaffenberger and Dielman derive size-corrected significance points from independent simulations pertaining to the particular regressor data sets considered.

Given the above, it seems clear that for the regression model, commonly tabulated critical points of standard normality tests can be quite misleading and should be improved. In this paper, we re-emphasize this fact and propose the use of the Monte Carlo (MC) test technique (Dwass, 1957; Barnard, 1963; Birnbaum, 1974; Jöckel, 1986; Dufour, 1995; Dufour and Kiviet, 1996, 1998; Kiviet and Dufour, 1997) in order to obtain finite sample *p*-values. In particular, we implement the procedures in Dufour (1995) relating to test statistics that are not necessarily continuous. This technique allows one to obtain exact (randomized) tests, in the sense that the probability of a type I error is known, whenever the null distribution of a test statistic does not depend on unknown parameters and can be simulated. Further, very small numbers of replications of the test statistics are required for that purpose. On observing that all standard normality test statistics are pivotal when applied to regression residuals, we suggest that MC testing provides an attractive alternative to usual asymptotic approximations. Indeed, the latter become irrelevant. Further, the proposed techniques can be extended easily to test other distributions (besides the normal), but we shall not stress this possibility here.

These finite sample properties hold whenever the regressor matrix is fixed or is random but independent of the disturbance vector (strict exogeneity). In the latter case, the results obtain through conditioning upon the regressor matrix. Even though this setup extends considerably earlier finite sample results in the area of testing normality (which are largely limited to testing the normality of *i.i.d.* observations), it is clear our regression model excludes many econometric setups, such as models with lagged dependent variables (dynamic models), weakly exogenous regressors or non *i.i.d.* disturbances (heteroskedasticity, serial correlation). However, it is worthwhile noting that the simulation-based procedure proposed here yield 'asymptotically valid' tests whenever the test criterion used has a nuisance-parameter-free null distribution under a class of data-generating processes which includes the (more restricted) ones considered here. For a related discussion, the reader may consult Dufour and Kiviet (1998).

MC tests are closely related to the parametric bootstrap, although with a fundamental difference. Whereas bootstrap tests are on the whole asymptotic (as the number of simulated samples goes to infinity), MC test methods yield provably exact tests, in the sense that the number of replications used is explicitly taken into account. Bootstrap methods have recently been suggested for GOF problems; see, for example, Stute *et al*. (1993) and Henze (1996). These authors present the bootstrap as an alternative asymptotic approach to treat empirical processes with estimated parameters. Monte Carlo studies were carried out for various parametric models with the conclusions that bootstrap Kolmogorov–Smirnov and Cramér–von Mises tests achieve level control. Although Stute *et al.* examined normality tests in the location-scale context as a special case, the problem has not apparently been considered from a finite sample perspective. Several authors have also advocated the use of the bootstrap for different (although related) specification tests in non-linear contexts; see, for example, Andrews (1997), Beran and Miller (1989) and Linton and Gozalo (1997). For further discussion of bootstrap methods, the reader may consult Efron (1982), Efron and Tibshirani (1993), Hall (1992), Jeong and Maddala (1993), Vinod (1993) and Shao and Tu (1995).

We also investigate the size and power of suggested tests in a Monte Carlo study across six error distributions. We consider several choices for the sample size, the number of regressors and the design matrix. In addition, we examine the effect on power of increasing the number of MC replications. The results show that MC tests overcome the usual size problems and achieve good power, even with small numbers of MC replications.

The paper is organized as follows. In Section 2, we set notation and review the test statistics under consideration. In Section 3, we discuss the pivotal character of the test statistics and present the MC test procedure. Section 4 reports the results of the simulation experiment. We conclude in Section 5.

2. MODEL AND TEST STATISTICS

We consider normality tests in the context of the linear regression model:

$$Y = X\beta + u, \tag{2.1}$$

where $Y = (y_1, \dots, y_n)'$ is a vector of observations on the dependent variable, X is the matrix of n observations on k regressors, β is a vector of unknown coefficients and u is an n-dimensional vector of *i.i.d.* disturbances; further, X is fixed or independent of u. The problem is to test

$$H_0 : f(u) = \varphi(u; 0, \sigma), \sigma > 0, \tag{2.2}$$

where $f(u)$ is the unknown probability density function (p.d.f.) and $\varphi(u; \mu, \sigma)$ is the normal p.d.f. with mean μ and standard deviation σ. The assumption that u has mean zero is not restrictive when X includes a constant term $\iota_n = (1, \dots, 1)'$. When $X = \iota_n$ the above regression model reduces to the location-scale model. In this context, we shall consider normality tests based on the least-squares residual vector

$$\hat{u} = y - X\hat{\beta} = M_X u, \tag{2.3}$$

where $\hat{\beta} = (X'X)^{-1}X'y$ and $M_X = I_n - X(X'X)^{-1}X'$. Let $\hat{u}_{1n} \le \hat{u}_{2n} \le \cdots \le \hat{u}_{nn}$ denote the order statistics of the residuals, and

$$s^2 = (n-k)^{-1} \sum_{i=1}^{n} \hat{u}_{in}^2, \quad \hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \hat{u}_{in}^2. \tag{2.4}$$

The tests we shall study can be grouped into three categories: empirical distribution function (EDF) tests, moment tests and correlation tests.

2.1. EDF tests

EDF tests are based on a measure of discrepancy between the empirical and hypothesized distributions. The most familiar EDF tests are: the Kolmogorov–Smirnov (KS) test (Kolmogorov, 1933; Smirnov, 1939), the Cramér–von Mises (VM) test (Cramér, 1928) and the Anderson–Darling (AD) test (Anderson and Darling, 1954). The finite sample distributions of the AD and VM statistics are quite complicated but an asymptotic theory is available. For the KS statistic, the exact and limiting distributions are non-standard and even asymptotic points must be estimated; this fact was first observed by Lilliefors (1967) who gave significance points by Monte Carlo calculations. To improve performance in finite samples, Stephens (1974) has proposed modifying the EDF statistics through multiplication by an appropriate correction factor; this author supplies adjustment formulas and approximate critical points for use with modified criteria. Revised significance points are also available in D'Agostino and Stephens (1986, Table 4.7). As pointed out above these pertain to the location-scale model.

The statistics are defined as follows:

$$KS = \max(D^+, D^-), \tag{2.5}$$

where $D^+ = \max_{1 \le i \le n}\{(i/n) - \hat{z}_i\}$ and $D^- = \max_{1 \le i \le n}\{\hat{z}_i - (i-1)/n\}$,

$$VM = \sum_{i=1}^{n} \{\hat{z}_i - (2i-1)/2n\}^2 + (1/12n), \tag{2.6}$$

$$AD = -n - n^{-1} \sum_{i=1}^{n} (2i - 1)\{\ln \hat{z}_i + \ln(1 - \hat{z}_{n+1-i})\}, \tag{2.7}$$

where $\hat{z}_i = \Phi(\hat{u}_{in}/s)$, $i = 1, \ldots, n$, and $\Phi(.)$ denotes the cumulative $N(0, 1)$ distribution function. In this article, we study both standard and modified (following Stephens) statistics; the modified statistics will be denoted KS_s, VM_s and AD_s.

2.2. Moment tests

Moment tests derive from the recognition that the third and fourth moments of the $N(0, 1)$ distribution are equal to 0 and 3, respectively. Hence, deviations from normality may be assessed using the sample moments, i.e. the coefficients of skewness (Sk) and kurtosis (Ku):

$$Sk = n^{-1} \sum_{i=1}^{n} \hat{u}_{in}^{3}/(\hat{\sigma}^2)^{3/2}, \; Ku = n^{-1} \sum_{i=1}^{n} \hat{u}_{in}^{4}/(\hat{\sigma}^2)^2. \tag{2.8}$$

The literature on the null distributions of these statistics and their joint density is vast. Although very few finite sample results are known, asymptotic theory is well developed and tables have been available for some time (see D'Agostino and Stephens (1986, Ch. 6)). The skewness and kurtosis tests may be implemented as two distinct tests. Procedures involving Sk and Ku jointly are also in common use. One popular example is the Jarque–Bera (JB) test (Jarque and Bera, 1980, 1987) based on a Lagrange multiplier criterion:

$$JB = n \left\{ \frac{1}{6}(Sk)^2 + \frac{1}{24}(Ku - 3)^2 \right\}. \tag{2.9}$$

As pointed out by Jarque and Bera (1987, p. 165), their method was independently suggested by Bowman and Shenton (1975) as an omnibus procedure combining Sk and Ku in one test statistic. Jarque and Bera have shown that the test derives from the LM principle in the context of the Pearson family of probability density functions. Under the null and appropriate regularity conditions, the JB statistic is asymptotically distributed as $\chi^2(2)$. As is typically the case with the various normality tests, the exact distribution is intractable. We have also considered moment tests where $\hat{\sigma}^2$ is replaced by s^2, which we denote Sk_k, Ku_k and JB_k, respectively.

2.3. Correlation tests

Correlation tests are based on the ratio of two estimates of scale obtained from order statistics: a weighted least-squares estimate given that the population is normally distributed and the unbiased estimate of scale for any population, i.e. the sample variance. The weights originally proposed for the Shapiro–Wilk (SW) test (Shapiro and Wilk, 1965) are the optimal weights in the sense of GLS estimation and are difficult to compute:

$$SW = \frac{(\sum_{i=1}^{n} a_i \hat{u}_{in})^2}{(n - k)s^2}, a' = (a_1, \ldots, a_n) = \frac{c'V^{-1}}{(c'V^{-2}c)^{1/2}} \tag{2.10}$$

where $c = (c_1, \ldots, c_n)'$ and V are respectively the vector of expected values and the covariance matrix of standard normal order statistics. Shapiro and Wilk (1965) supply a table of weights and significance points for location-scale models with $n \leq 50$; these are reproduced in D'Agostino and Stephens (1986, Tables 5.4 and 5.5). For large samples, Shapiro and Francia (1972) suggest ignoring the covariance term in the formulae for deriving the weights; in other words, the Shapiro–Francia (SF) test treats the ordered observations as if they were independent:

$$SF = \frac{(\sum_{i=1}^{n} b_i \hat{u}_{in})^2}{(n-k)s^2}, b' = (b_1, \ldots, b_n) = \frac{c'}{(c'c)^{1/2}}. \tag{2.11}$$

The SF statistic may also be interpreted as the correlation coefficient between c and the order statistics of the residuals. Shapiro and Francia supplied the weights and significance points for location-scale models with $n < 100$; D'Agostino and Stephens (1986, Table 5.2) provides the critical values of $n(1 - SF)$ for location-scale models with $n \leq 1000$. Royston (1982a, b, c) has also published algorithms for computing the distribution of the SW statistic, but these only apply to simple location-scale models.

D'Agostino (1971) proposed considering a linear combination of the ordered observations that does not require a table of weights. The D'Agostino (D) statistic may be computed as follows:

$$D = \frac{\sum_{i=1}^{n} \hat{u}_{in}\{i - (n+1)/2\}}{n^{3/2}\{(n-k)s^2\}^{1/2}}. \tag{2.12}$$

D'Agostino (1971, 1972) provide significance points for location-scale models with $n \leq 2000$; these are reproduced in D'Agostino and Stephens (1986, Table 9.7). Several other modified SF statistics have been suggested. We consider the Weisberg–Bingham (WB) test (Weisberg and Bingham, 1975) and the Filliben (FB) test (Filliben, 1975). The WB statistic derives from the SF statistic substituting the following for *c:*

$$\hat{c}_i = \Phi^{-1}\left\{\frac{i - (3/8)}{n + (1/4)}\right\}, i = 1, \ldots, n, \tag{2.13}$$

where Φ^{-1} refers to the inverse of the standard normal cumulative distribution function. The critical values of the test are those of the SF test. The FB criterion may be viewed as the correlation coefficient between the ordered residuals and the order statistics medians from the standard normal distribution. Filliben produced weights and critical points for the location-scale model with $n \leq 100$.

3. MONTE CARLO TESTS FOR NORMALITY

All of the existing tables of critical points described above were generally derived from Monte Carlo simulations following the *i.i.d.* location-scale model. As an alternative to these, we shall employ the technique of MC tests. To provide necessary background, we first discuss relevant invariance properties of the statistics considered. The MC test procedure is described next.

3.1. Pivotal property of standardized residuals

From (2.5) to (2.12), we see that all the test statistics can be computed from the standardized residual vector $\hat{u}/s$. Using (2.3), we can write:

$$\hat{u}/s = \frac{\hat{u}}{(\hat{u}'\hat{u}/(n-k))^{1/2}} = (n-k)^{1/2}\frac{M_X u}{(u' M_X u)^{1/2}} = (n-k)^{1/2}\frac{M_X w}{(w' M_X w)^{1/2}}, \tag{3.14}$$

where the components of $w = u/\sigma$ are *i.i.d.* $N(0, 1)$ when $u \sim N(0, \sigma^2 I_n)$, so that $\hat{u}/s$ follows a nuisance-parameter free distribution. The distribution of the scaled vector $\hat{u}/s$ depends on the (known) regressor matrix X, but not on the regression parameters β and σ. When X is fixed, this entails that $\hat{u}/s$ follows a nuisance-parameter-free distribution. When X is viewed as random but remains independent of u, the marginal distribution of $\hat{u}/s$ may depend on the parameters of the distribution of X, but its conditional distribution given X only depends on X. Consequently, in both situations, residual-based test statistics are location and scale invariant, and their exact null distributions can be simulated easily.

3.2. Monte Carlo test procedure

Let T be a real-valued test statistic such that a null hypothesis of interest H_0, e.g. model (2.1) with $u \sim N(0, \sigma^2 I_n)$, is rejected when T is large, i.e. when $T \geq c$, where the constant c depends on the level of the test, and suppose T is pivotal. In other words, given a statistical model $(\Omega, \mathcal{A}, \mathcal{P})$ where Ω is a sample space, $\mathcal{A}$ is a $\sigma-$algebra of subsets of Ω and $\mathcal{P}$ is a family of probability measures on $\mathcal{A}$ which include the set $\mathcal{P}_0 \subseteq \mathcal{P}$ of measures compatible with H_0, we assume $T = T(\omega)$ is a mapping from Ω to $\mathbb{R}$ ($T : \Omega \to \mathbb{R}$) such that the survival function $G(x) \equiv P(T \geq x) \equiv P[\{\omega \in \Omega : T(\omega) \geq x\}]$, or equivalently the distribution function $F(x) = P(T \leq x)$, is the same for all $P \in \mathcal{P}_0$ (so that the critical region $T \geq c$ is similar). Note the function $G : \mathbb{R} \to [0, 1]$ does not depend on ω and must be viewed as fixed (hence independent of any random variable defined on Ω) in the present context. Then $G(c) = \alpha$ is the size of the critical region $T \geq c$. Further, for any $\mathcal{A}$-measurable random variable $T_0 = T_0(\omega_0)$, $\omega_0 \in \Omega$, the transformed random variable $G(T_0) \equiv G\{T_0(\omega_0)\}$, $\omega_0 \in \Omega$, satisfies $P\{G(T_0) \leq \alpha\} = \alpha$, where $P\{G(T_0) \leq x\} \equiv P[\omega_0 \in \Omega : G\{T(\omega_0)\} \leq x]$ for any $x \in \mathbb{R}$. Note the random variable $G(T_0)$ can be interpreted as the conditional probability $P(T \geq T_0 | T_0)$ when T and T_0 are *i.i.d.* (defined on the appropriate product measure space) each with the survival function $G(x)$; further, if T_0 denotes the test statistic computed from data (a random variable) and $\mathcal{T}_0$ the observed value of T_0 based on specific realized data (taken as given (fixed)), $G(\mathcal{T}_0) = P(T \geq T_0 | T_0 = \mathcal{T}_0)$ is the 'realized' p-value of the test statistic T_0.

Now suppose we can generate N independent realizations $T_1, \ldots, T_N$, from which we can compute an empirical p-value function:

$$\hat{p}_N(x) = \frac{N\hat{G}_N(x) + 1}{N + 1} \tag{3.15}$$

where

$$\hat{G}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[0,\infty)}(T_i - x), \mathbf{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} . \tag{3.16}$$

The associated MC critical region is a randomized critical region defined as

$$\hat{p}_N(T_0) \leq \alpha \tag{3.17}$$

where $\hat{p}_N(T_0)$ may be interpreted as an estimate of $G(T_0)$. When the distribution of T_0 is continuous, we have:

$$P\{\hat{p}_N(T_0) \leq \alpha\} = \frac{I\{\alpha(N + 1)\}}{N + 1}, \text{ for } 0 \leq \alpha \leq 1, \tag{3.18}$$

where $I[x]$ denotes the largest integer less than or equal to x; see Dufour (1995), Dufour and Kiviet (1996, 1998) or Kiviet and Dufour (1997). Given $T_0 = \mathcal{T}_0$, the quantity $\hat{p}_N(\mathcal{T}_0)$ may be interpreted as a (randomized) 'realized' p-value associated with T_0. Thus, if N is chosen such that $\alpha(N+1)$ is an integer, the critical region (3.17) has the same size as the critical region $G(T_0) \leq \alpha$. The MC test so obtained is theoretically exact, irrespective of the number N of replications used.

The above procedure is closely related to a parametric bootstrap, but with a fundamental difference. Bootstrap tests are, in general, provably valid for $N \to \infty$. In contrast, we see from (3.18) that N is explicitly taken into consideration in establishing the validity of MC tests. Although the value of N has no incidence on size control, it may have an impact on power which typically increases with N.

Note that (3.18) holds for tests based on statistics with continuous distributions. In the case of the KS criterion, ties have non-zero probability. Nevertheless, the technique of MC tests can be adapted for discrete distributions by appeal to the following randomized tie-breaking procedure (see Dufour (1995)).

Draw $N+1$ uniformly distributed variates $W_0, W_1, \ldots, W_N$, independently of T_j and arrange the pairs (T_j, W_j) following the lexicographic order:

$$(T_i, W_i) \geq (T_j, W_j) \Leftrightarrow \{T_i > T_j \text{ or } (T_i = T_j \text{ and } W_i \geq W_j)\}. \tag{3.19}$$

Then, proceed as in the continuous case and compute

$$\tilde{p}_N(x) = \frac{N\tilde{G}_N(x) + 1}{N+1}, \tag{3.20}$$

where

$$\tilde{G}_N(x) = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[0,\infty)}(x - T_i) + \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{[0]}(T_i - x)\mathbf{1}_{[0,\infty)}(W_i - W_0). \tag{3.21}$$

The resulting critical region$\tilde{p}_N(T_0) \leq \alpha$ has the same level as the region $G(T_0) \leq \alpha$, again provided $\alpha(N+1)$ is an integer. More precisely,

$$P\{\hat{p}_N(T_0) \leq \alpha\} \leq P\{\tilde{p}_N(T_0) \leq \alpha\} = \frac{I\{\alpha(N+1)\}}{N+1}, \text{ for } 0 \leq \alpha \leq 1.$$

If $T_0, T_1, \ldots, T_N$ are all distinct, $\tilde{p}_N(T_0) = \hat{p}_N(T_0)$.

The procedures discussed in this section can be readily extended to other GOF hypotheses. Indeed, the central properties we have exploited here are the following: (i) the standardized error vector has a known null distribution, and (ii) the test statistics depend only on the empirical distribution function of residuals. These properties are preserved for: (i) all error distribution functions which are completely specified up to a scale parameter, and (ii) any relevant GOF criterion based on the empirical process of residuals. The latter generalization allows for a natural class of GOF statistics, although others may be worth consideration. Of course, the choice of which statistic to employ depends on the specific hypothesis at hand.

4. SIMULATION EXPERIMENT

The simulation experiment was performed as follows. The model used was (2.1). For each disturbance distribution, the tests were applied to the residual vector, obtained as $\hat{u} = M_X u$.

Table 1.1. List of abbreviations

Notation	Test	Reference
KS	Kolmogorov–Smirnov	Equation (2.5)
KS_s	Modified KS	D'Agostino and Stephens (1986, Table 4.7)
VM	Cramér–von Mises	Equation (2.6)
VMs	Modified VM	D'Agostino and Stephens (1986, Table 4.7)
AD	Anderson–Darling	Equation (2.7)
ADs	Modified AD	D'Agostino and Stephens (1986, Table 4.7)
JB	Jarque–Bera	Equations (2.9) and (2.4)
JB_k	Jarque–Bera (using s^2)	Equations (2.9) and (2.4)
SW	Shapiro–Wilk	Equation (2.10)
SF	Shapiro–Francia	Equation (2.11)
WB	Weisberg–Bingham	Equations (2.11) and (2.13)
D	D'Agostino	Equation (2.12)
FB	Filliben	Filliben (1975)

Table 1.2. Critical points for standard normality tests

Test	Reference	Sample size			
		25	50	100	300
KS	Lilliefors (1967)	0.173	$0.886/\sqrt{n}$	$0.886/\sqrt{n}$	$0.886/\sqrt{n}$
VM	D'Agostino and Stephens (1986, Table 4.10)	0.12125	0.1225	0.125	0.126
AD	D'Agostino and Stephens (1986, Table 4.10)	0.71625	0.7285	0.742	0.752
KS_s	D'Agostino and Stephens (1986, Table 4.7)	0.895	0.895	0.895	0.895
VM_s	D'Agostino and Stephens (1986, Table 4.7)	0.126	0.126	0.126	0.126
AD_s	D'Agostino and Stephens (1986, Table 4.7)	0.752	0.752	0.752	0.752
SF	D'Agostino and Stephens (1986, Table 5.2)	1.99	2.31	2.56	2.67
WB	D'Agostino and Stephens (1986, Table 5.2)	1.99	2.31	2.56	2.67
SW	D'Agostino and Stephens (1986, Table 5.5)	0.918	0.947	n.a.	n.a.
D	D'Agostino and Stephens (1986, Table 9.7)	−2.97	−2.74	−2.54	−2.316
		0.74	1.06	1.31	1.528
FB	Filliben (1975)	0.958	0.977	0.987	n.a.

In the presented table, asterisks indicate the highest computed power achieved in each column. The modified EDF statistics KS_S, VM_S and AD_S are monotonic transformations of the original criteria KS, VM and AD, respectively, and so yield the same MC p-values. The result for tests based on standard critical values may not be reported in a few cases (like SW and FB in Table 3) because the required critical values have not apparently been tabulated for the regression design considered.

Hence, there was no need to specify the coefficient vector β. The matrix X included a constant term, a set of k_1 dummy variables and a set of independent standard normal variates. Formally,

$$X = \left\{ \iota_n \vdots \quad X_{(1)} \vdots \quad X_{(2)} \right\}, X_{(1)} = \left\{ \begin{array}{l} I_{k_1} \\ 0_{(n-k_1,k_1)} \end{array} \right\} \tag{4.22}$$

Table 2. Empirical size and power of normality tests; *i.i.d.* observations

	Standard Tests						MC Tests					
	N	B	C	Γ	Ln	t	N	B	C	Γ	Ln	t
$n = 25$												
KS	5.3	7.2	90.2	39.6	98.5	14.8	5.2	7.3	90.0	38.5	98.4	14.6
KS_s	5.2	7.0	90.5	39.3	98.5	14.7	5.2	7.3	90.0	38.5	98.4	14.6
VM	5.6	8.7	93.6	51.6	99.7	18.6	5.1	8.3	93.0	49.4	99.6	17.2
VM_s	5.2	8.1	93.3	50.4	99.7	17.9	5.1	8.3	93.0	49.4	99.6	17.2
AD	5.6	9.1	93.7	57.5	99.9	20.2	5.2	8.6	93.0	54.8	99.8	19.2
AD_s	5.1	8.3	93.4	55.8	99.9	19.4	5.2	8.6	93.0	54.8	99.8	19.2
JB	2.9	0.9	89.5	37.8	95.9	21.2	5.2	2.1	91.4	47.8	97.5	26.3*
JB_k	1.8	0.4	87.2	31.6	94.1	17.4	5.2	3.0	91.0	51.2	98.2	25.6
SW	5.2	8.4	92.2	64.2	100	21.1	5.4	8.7*	92.0	63.3*	99.9*	21.3
SF	5.8	5.1	94.0	61.5	99.9	26.5	5.2	4.6	93.7	58.3	99.8	25.5
WB	5.7	5.1	94.0	61.5	99.9	26.4	5.3	4.7	93.7	58.4	99.8	25.1
D	5.4	7.1	93.4	33.1	97.3	22.4	5.2	7.0	92.6	30.7	96.5	21.3
FB	5.3	4.4	94.0	59.4	99.9	26.1	5.2	4.3	93.8*	57.7	99.8	25.4
$n = 50$												
KS	4.6	11.5	99.4	68.0	100	20.5	4.9	11.7	99.3	67.7	100*	20.8
KS_s	4.8	12.0	99.4	69.1	100	21.2	4.9	11.7	99.3	67.7	100*	20.8
VM	5.4	15.5	99.7	83.7	100	27.8	5.0	14.7	99.7	81.8	100*	26.8
VM_s	5.1	14.9	99.7	83.0	100	27.2	5.0	14.7	99.7	81.8	100*	26.8
AD	5.3	18.1	99.7	89.1	100	31.0	5.0	16.9	99.7	87.6	100 *	29.8
AD_s	5.0	17.2	99.7	88.5	100	30.2	5.0	16.9	99.7	87.6	100*	29.8
JB	3.7	0.8	99.5	76.0	100	39.4	4.8	3.0	99.5	79.8	99	41.8*
JB_k	2.7	0.5	99.4	72.7	100	36.1	4.9	4.9	99.5	82.9	100*	41.0
SW	4.3	26.2	99.4	94.8	100	26.4	5.0	27.8*	99.4	94.8 *	100*	27.2
SF	5.1	10.3	99.8	91.9	100	41.3	5.0	10.0	99.8*	90.9	99	40.5
WB	5.1	10.1	99.8	91.8	100	41.5	5.0	9.9	99.8*	90.8	100*	40.5
D	5.3	13.7	99.8	56.4	100	39.0	5.2	13.1	99.7	53.6	100*	36.9
FB	5.6	9.9	99.8	91.9	100	43.4	5.0	8.8	99.7	90.0	100*	41.1

(cont.)

where $0_{(i,j)}$ denotes an (i, j) matrix of zeros, $X_{(2)}$ includes $k - k_1 - 1$ regressors drawn as *i.i.d.* standard normal. Sample sizes of $n = 25$, 50, 100 (and 300 in certain cases) were used, k was set as the largest integer less than or equal to $\sqrt{n}$ and $k_1 = 0, 2, 4, \ldots, k - 1$. We have also examined the cases where (i) $X = \iota_n$, i.e. the location-scale model, and (ii) X includes a constant term and $k - 1$ regressors drawn from a Cauchy distribution. As mentioned earlier, the regressors here are treated as fixed across replications, which excludes many cases of interest in econometrics such as lagged dependent variables (dynamic models).

Table 2. Continued

	Standard Tests						MC Tests					
	N	B	C	Γ	Ln	t	N	B	C	Γ	Ln	t
$n = 100$												
KS	5.0	23.5	100	95.4	100	33.5	4.8	22.7	100*	94.5	100*	31.9
KS_s	4.9	23.2	100	95.2	100	33.1	4.8	22.7	100*	94.5	100*	31.9
VM	4.8	32.0	100	99.1	100	42.8	4.9	31.4	100*	99.0	100*	42.1
VM_s	4.9	32.3	100	99.1	100	43.1	4.9	31.4	100*	99.0	100*	42.1
AD	5.0	40.7	100	99.8	100	48.1	4.8	39.1*	100*	99.7	100*	47.1
AD_s	4.9	40.0	100	99.8	100	47.9	4.8	39.1*	100*	99.7	100*	47.1
JB	3.9	4.7	100	99.1	100	62.8	5.0	12.6	100*	98.6	100*	63.8
JB_k	3.4	5.2	100	99.0	100	60.2	4.9	19.2	100*	99.1	100*	62.9
SF	4.6	33.3	100	99.9	100	61.7	4.8	32.6	100*	99.9*	100*	61.4
WB	4.7	32.4	100	99.9	100	62.2	4.8	31.4	100*	99.9*	100*	61.6
D	5.1	29.0	100	82.5	100	62.9	5.2	26.3	100*	79.8	100*	60.5
FB	4.9	29.1	100	99.9	100	63.1	4.8	28.0	100*	99.9 *	100*	62.1

The disturbances were generated from several distributions: standard normal, Cauchy, lognormal, beta(2,3), gamma(2,1) (denoted N, C, LN, B, Γ respectively) and Student $t(5)$. We assessed the performance of all the tests reviewed above at the nominal size of 5%. With the exception of the D test, all were treated as one-sided tests; the relevant critical points for the standard tests are given in Table 1.2. Tables 2 to 5 report the rejection percentages among 10 000 replications.

The MC procedures illustrated in Tables 2 to 4 are based on 99 simulated samples (79 in the case of the D statistic). We have also examined the effect on power of increasing the number of simulated samples. Results for these experiments are presented in Table 5, where $N = 19, 29, \ldots, 99, 199, \ldots, 499$. For the D statistic, N was set to $39, 79, 199$ and 399. In the presented tables, asterisks indicate the highest computed power achieved in each column. The modified EDF statistics KS_s,VM_s and AD_s are monotonic transformations of the original criteria KS, VM and AD, respectively, and so yield the same MC p-values. The results for tests based on standard critical values may not be reported in a few cases (like SW and FB in Table 3) because the required critical values have not apparently been tabulated for the regression design considered. More complete results (with graphs) are available in a technical report (Dufour *et al.*, 1997). Our conclusions may be summarized as follows.

4.1. Test size

The location-scale model. For the simple location-scale model, all the tests except the JB procedure control size reasonably well (see Table 2). The EDF, the SW and the D tests appear adequate. While the SF, WB and FB tests tend to over-reject, the distortions are not severe.

Table 3. Empirical size of normality tests based on regression residuals

	$n = 25,\ k = 5$						$n = 50,\ k = 7$							
	Standard tests			MC tests			Standard tests				MC tests			
k_1:	0	2	4	0	2	4	0	2	4	6	0	2	4	6
KS	5.2	11.6	28.6	5.2	5.3	5.2	5.3	7.9	15.7	29.4	5.2	5.0	5.1	5.0
KS_s	5.0	11.4	28.2	5.2	5.3	5.2	5.5	8.1	16.3	30.1	5.2	5.0	5.1	5.0
VM	6.8	12.7	28.3	5.2	5.2	5.0	7.4	10.5	18.9	33.9	5.1	5.0	4.9	4.9
VM_s	6.4	12.0	27.1	5.2	5.2	5.0	6.9	9.9	17.9	32.5	5.1	5.0	4.9	4.9
AD	6.4	10.5	22.0	5.2	5.2	5.3	7.2	9.5	15.9	26.9	5.2	5.0	4.8	4.8
AD_s	5.7	9.7	20.7	5.2	5.2	5.3	6.7	8.9	14.9	25.7	5.2	5.0	4.8	4.8
SF	5.7	8.4	14.6	5.3	5.2	5.0	5.2	6.5	9.2	13.7	5.0	5.3	5.2	4.8
SW	5.1	6.2	10.2	5.5	5.3	5.2	4.2	4.1	5.0	6.9	4.9	5.0	5.0	4.8
WB	5.7	8.4	14.5	5.4	5.2	5.0	5.2	6.5	9.2	13.8	5.0	5.3	5.2	4.8
D	5.0	6.6	11.4	5.0	5.3	5.1	5.1	5.7	7.6	12.4	5.0	5.1	4.8	5.2
FB	5.2	7.9	13.9	5.4	5.1	5.1	5.7	7.3	10.3	15.3	5.1	5.3	5.2	4.7
JB	2.9	4.8	6.7	5.2	5.2	4.8	3.9	5.1	6.4	8.4	5.0	5.1	5.0	4.7
JB_k	0.1	0.2	0.4	5.1	5.0	5.1	0.3	0.5	0.8	1.0	4.8	5.1	4.9	5.0

	$n = 100, k = 11$											
	Standard tests						MC tests					
k_1	0	2	4	6	8	10	0	2	4	6	8	10
KS	5.6	8.0	13.3	20.1	32.6	47.6	4.7	5.0	5.0	4.7	4.7	5.0
KS_s	5.5	7.9	13.1	19.7	32.2	47.2	4.7	5.0	5.0	4.7	4.7	5.0
VM	7.4	9.6	15.0	21.8	33.1	48.8	4.9	4.8	4.8	4.9	4.8	5.0
VM_s	7.6	9.7	15.1	22.2	33.4	49.3	4.9	4.8	4.8	4.9	4.8	5.0
AD	7.8	9.8	13.5	18.9	27.9	39.9	4.8	4.7	5.0	4.9	4.7	4.9
AD_s	7.6	9.5	13.2	18.5	27.6	39.4	4.8	4.7	5.0	4.9	4.7	4.9
SF	4.8	5.1	6.6	8.6	11.2	15.4	5.0	4.5	4.9	5.0	4.6	4.8
WB	4.9	5.2	6.8	8.8	11.4	15.7	5.0	4.5	4.9	5.0	4.7	4.9
D	5.1	5.3	6.3	7.9	10.8	15.4	5.3	4.8	5.1	5.1	4.9	5.1
FB	4.9	5.3	7.0	9.0	11.9	16.2	5.0	4.5	4.9	5.1	4.7	4.8
JB	4.1	4.7	5.8	7.1	8.9	10.2	4.8	4.8	4.8	4.9	4.9	5.1
JB_k	2.1	1.6	1.5	1.4	1.3	1.5	5.2	4.7	5.0	5.0	4.8	4.7

(cont.)

However, the JB test substantially under-rejects. The sizes of all MC tests correspond closely to the nominal value of 5%.

The regression model. From the results in Table 3, we can see that the test performances in the regression context can be much worse than for the location-scale model. Although the tests appear adequate when the explanatory variables are generated as standard normal, the sizes of all

Table 3. Continued

	$n = 300, k = 17$									Cauchy regressors			
	Standard tests									Standard tests: $(n, k) =$			
k_1	0	2	4	6	8	10	12	14	16	(25, 6)	(50, 8)	(100, 11)	(300, 17)
KS	6.7	8.1	9.9	12.7	16.6	21.8	27.9	35.2	43.9	10.9	11.3	11.1	10.7
KS_s	6.2	7.6	9.3	12.0	15.9	20.8	26.6	34.0	42.7	10.7	11.8	10.8	10.1
VM	7.0	8.3	10.3	12.9	16.1	20.5	26.9	33.5	42.3	14.8	15.3	14.7	12.5
VM_s	7.0	8.9	10.4	13.0	16.2	20.6	27.1	33.6	42.4	13.8	14.5	14.5	12.5
AD	7.4	8.5	9.9	12.0	14.5	17.9	22.4	28.0	34.4	12.9	13.6	14.4	12.5
AD_s	7.5	7.1	10.0	12.1	14.6	18.1	22.5	28.2	34.6	11.9	12.9	14.2	12.6
SF	6.3	7.3	7.2	8.3	8.4	10.1	11.8	13.5	16.2	10.3	9.1	7.9	8.8
SW	—	—	—	—	—	—	—	—	—	7.3	5.0	—	—
WB	6.4	5.1	7.5	8.5	8.6	10.4	12.1	13.8	16.6	10.3	9.2	8.2	9.1
D	4.6	5.3	5.2	5.7	6.7	7.6	9.2	10.8	13.2	8.3	7.5	7.5	6.5
FB	—	—	—	—	—	—	—	—	—	9.9	10.2	8.5	—
JB	4.5	5.1	5.5	5.8	6.5	7.1	8.2	8.8	9.8	6.1	7.3	7.8	7.6
JB_k	8.0	7.1	6.0	5.3	4.9	4.5	4.1	3.8	3.7	0.3	0.8	1.5	5.1

tests vary substantially from the nominal 5% for all other designs, irrespective of the sample size. More specifically, (i) the EDF tests consistently over-reject and the modified versions over-reject by the same magnitude, (ii) the correlation tests over-reject but to a lesser extent, (iii) the moment tests based on s are severely undersized and, (iv) the moment tests based on $\hat{\sigma}$ under-reject when the number of dummy variables relative to normal regressors is small and over-reject otherwise. In contrast, all MC tests achieve perfect size control for all sample sizes.

An interesting experiment that bears on this problem is reported in Weisberg (1980). Weisberg had pointed out that in the context of normality tests, Monte Carlo results based on data sets where all explanatory variables are drawn from the uniform or standard normal distribution are not representative and that size problems may occur. He demonstrated this with a specific data set for the SW test with $n = 20$. The analysis here extends this observation in two important ways. First, we show that problems can occur for *all* conventional tests. Second, the design matrices we consider involve samples as large as 100 and 300 and are quite likely to be encountered in econometric practice. An intuitive explanation for the effect of dummy variables on test size is the following. Residuals based on normal regressors may mimic an *i.i.d.* series if k is small enough, relative to n. The appended indicator variables cause k_1 residuals to be zero, and these should be excluded from the test procedure (but are not). This provides a simple example where standard distributional theory fails, while our approach works without any difficulty. Regressors drawn form a Cauchy distribution (see Table 3) provide another although less extreme example of such situations. Note, finally, that in Table 3 the level does not appear to be better controlled as the sample size increases. This is simply due to the fact that, in this experiment, the number of regressors increases with sample size.

Table 4. Empirical power of MC normality tests based on regression residuals

		B	C	Γ	Ln	t		B	C	Γ	Ln	t
KS	$n = 25$	4.5	74.3	22.6	79.8	12.7	$n = 25$	3.6	80.8	24.6	84.1	14.9
VM	$k = 5$	4.2	81.0	28.6	87.4	15.2	$k = 5$	2.9	86.6	28.4	91.0	18.1
AD	$k_1 = 0$	4.5	82.7	31.7	89.3	16.6	$k_1 = 2$	3.3	87.4	33.3	92.9	19.5
SF		3.8	83.7	34.9	90.7*	18.6		3.0	87.1	37.5	94.0	21.0
SW		6.0	80.1	35.8*	90.7*	15.4		4.8	85.3	42.1*	95.1*	19.0
WB		3.8	83.7	34.9	90.7*	18.5		3.0	87.1	37.6	94.0	21.0
D		5.3	81.3	19.9	81.4	15.3		6.4	84.0	21.4	84.2	16.2
FB		3.6	84.0*	34.4	90.6	18.8		2.9	87.3*	37.0	93.8	21.2
JB		2.5	83.5	32.3	88.8	19.9*		2.1	85.2	34.0	90.2	21.6*
JB_k		9.8*	69.4	20.4	80.3	9.8		9.8*	76.1	32.4	91.0	13.6
KS	$n = 25$	4.0	86.6	27.7	90.9	14.5						
VM	$k = 5$	2.1	91.5	26.7	94.6	20.1						
AD	$k_1 = 4$	2.2	91.8	33.0	96.6	21.6						
SF		2.4	91.5	40.5	97.7	23.6*						
SW		3.6	91.0	46.4*	98.7*	22.2						
WB		2.4	91.5	40.5	97.7	23.6						
D		7.5	88.4	22.3	88.8	17.9						
FB		2.2	91.6*	39.9	97.6	23.6						
JB		1.9	87.9	36.8	93.0	23.4						
JB_k		9.0*	82.7	43.5	97.8	18.1						
KS	$n = 50$	6.2	96.9	47.2	99.0	20.4	$n = 50$	4.5	97.7	47.6	99.2	21.6
VM	$k = 7$	5.9	98.6	59.6	99.8	26.1	$k = 7$	3.9	98.9	59.4	99.8	28.7
AD	$k_1 = 0$	6.7	98.8*	65.9	99.8	28.9	$k_1 = 2$	4.6	99.1	67.1	99.9*	31.4
SF		5.9	98.7	71.5	99.9*	33.1		4.8	99.0	73.8	99.9*	34.8
SW		15.2	97.0	73.2*	99.9*	21.6		15.0	98.1	78.5*	99.9*	25.1
WB		5.9	98.7	71.3	99.9*	33.2		4.7	99.0	73.8	99.9*	34.8
D		8.8	98.6	39.1	98.8	29.1		10.0	98.9	40.6	99.9*	30.1
FB		5.3	98.8*	70.7	99.9*	33.7		4.3	99.1*	73.0	99.9*	35.1
JB		2.1	98.6	64.1	99.7	35.0*		1.7	98.7	65.5	99.8	35.8*
JB_k		18.3*	95.4	59.9	99.5	19.9		19.2*	96.6	69.1	98.8	23.3

(cont.)

4.2. Test Power

The location-scale model. It is evident from Table 2 that MC tests correct for size and achieve good power. Overall, we do not observe any significant power loss for tests having comparable size. When interpreting the power of the correlation tests, keep in mind that the standard SF,

Table 4. Continued

		B	C	Γ	Ln	t		B	C	Γ	Ln	t
KS	$n = 50$	4.5	98.2	48.9	99.6	20.8	$n = 50$	5.2	98.9	52.8	99.8	20.0
VM	$k = 7$	2.8	99.3	56.2	99.8	31.2	$k = 7$	2.1	99.6	53.8	100*	32.4
AD	$k_1 = 4$	3.5	99.4*	65.8	99.9	34.0	$k_1 = 6$	2.6	99.7*	64.9	100*	36.1
SF		3.6	99.4*	75.9	99.9	38.0		2.8	99.6	77.5	100*	40.6
SW		12.8	98.9	83.1*	100*	29.8		11.9	99.4	87.3*	100*	33.4
WB		3.5	99.3	75.7	99.9	38.1		2.7	99.6	77.3	100*	40.6
D		10.8	99.2	40.3	99.3	31.3		13.1	99.4	40.9	99.5	32.9
FB		3.1	99.4*	74.9	99.9	38.4*		2.6	99.6	76.4	100*	40.8*
JB		1.4	98.9	65.9	99.8	38.2		1.1	99.1	66.6	99.8	40.0
JB_k		18.6*	97.4	77.3	99.9	27.9		17.3*	98.3	82.8	100*	32.4
KS	$n = 100$	9.7	100*	80.6	100*	33.1	$n = 100$	7.4	100*	80.2	100*	34.0
VM	$k = 11$	9.8	100*	91.1	100*	43.8	$k = 11$	7.2	100*	90.8	100*	46.1
AD	$k_1 = 0$	12.5	100*	94.6	100*	48.3	$k_1 = 2$	10.0	100*	94.6	100*	50.7
SF		16.0	100*	97.2*	100*	52.7		14.0	100*	97.5*	100*	55.4
WB		15.4	100*	97.2*	100*	52.9		13.5	100*	97.5*	100*	55.5
D		17.1	100*	66.6	100*	50.7		19.5	100*	67.2	100*	51.9
FB		13.6	100*	96.9	100*	53.9		12.0	100*	97.2	100*	56.0
JB		6.7	100*	94.1	100*	55.3*		4.6	100*	93.9	100*	57.0*
JB_k		42.1*	99.9	93.8	100*	32.9		45.1*	99.9	95.8	100*	36.2
KS	$n = 100$	6.8	100*	80.2	100*	33.2	$n = 100$	6.7	100*	80.4	100*	31.3
VM	$k = 11$	5.0	100*	90.1	100*	48.0	$k = 11$	3.7	100*	88.5	100*	48.9
AD	$k_1 = 4$	7.2	100*	94.6	100*	52.6	$k_1 = 6$	5.5	100*	94.3	100*	53.9
SF		11.7	100*	97.8*	100*	57.6		9.9	100*	98.0	100*	59.1
WB		11.3	100*	97.8*	100*	57.8		9.6	100*	98.0	100*	59.3
D		21.0	100*	67.9	100*	52.8		22.7	100*	67.7	100*	54.2
FB		9.8	100*	97.6	100*	58.3		8.3	100*	97.7	100*	59.3
JB		3.0	100*	93.8	100*	58.4*		2.0	100*	93.5	100*	59.7*
JB_k		46.8*	99.9	97.5	100*	39.7		48.4*	100*	98.2*	100*	43.1

WB and the FB tests are slightly oversized. Note also that the modified EDF (i.e. size corrected following Stephens (1974)) and the MC tests demonstrate similar power for all sample sizes across all the distributions examined. Most important is the effect of the MC procedure on the moment tests. Indeed, the effective power of the JB tests improves appreciably for $n \leq 100$. This is expected since the standard JB test is severely undersized.

The powers of the MC tests are broadly in the following order. The SW (when feasible) and the SF approximations are among the most powerful against practically all alternatives. The WB seems a sensible choice for it does not rely on any table of weights. However, for $n = 25$, the SF,

Table 5. The effect of the number of Monte Carlo replications on power

		Location-scale model, $n = 25$								
MC reps.		19	39	59	79	99	199	299	399	499
KS	B	7.2	7.3	7.4	7.4	7.3	7.3	7.2	7.2	7.2
	C	88.1	89.3	89.8	90.0	90.5	90.5	90.5	90.6	90.6
	Γ	34.0	36.9	37.4	38.2	38.5	38.7	39.0	39.2	39.2
	Ln	96.8	97.9	98.1	98.3	98.4	98.5	98.5	98.6	98.5
	t	13.6	14.4	14.5	14.6	14.6	14.8	14.7	14.7	14.8
VM	B	8.2	8.2	8.3	8.2	8.3	8.3	8.2	8.4	8.4
	C	91.7	92.6	93.0	93.0	93.0	93.2	93.2	93.3	93.3
	Γ	44.7	47.6	48.4	49.1	49.4	49.8	48.9	50.1	50.1
	Ln	99.1	99.5	99.5	99.6	99.6	99.7	99.7	99.7	99.7
	t	16.3	17.1	17.1	17.3	17.2	17.4	17.8	17.8	17.7
AD	B	8.7	8.5	8.7	8.6	8.6	8.7	8.6	8.7	8.7
	C	91.8	92.5	93.0	93.1	93.0	93.4	93.3	93.4	93.4
	Γ	49.9	52.9	54.0	54.7	54.8	55.5	55.8	56.0	55.9
	Ln	99.5	99.8	99.8	99.8	99.8	99.8	99.9	99.9	99.9
	t	17.8	18.8	19.0	19.2	19.2	19.5	19.6	19.7	19.7
SF	B	5.0	4.9	4.7	4.7	4.6	4.5	4.6	4.6	4.7
	C	92.3	93.3	93.5	93.6	93.7	93.7	93.8	93.8	93.8
	Γ	52.9	55.9	57.3	57.9	58.3	59.5	59.5	59.7	59.8
	Ln	99.4	99.7	99.8	99.8	99.9	99.9	99.9	99.9	99.9
	t	23.0	24.3	24.7	24.9	25.1	25.3	25.5	25.7	25.7
WB	B	5.0	4.9	4.7	4.7	4.7	4.6	4.6	4.6	4.7
	C	92.4	93.2	93.5	93.6	93.7	93.7	93.8	93.8	93.8
	Γ	53.0	56.0	57.4	58.0	58.4	59.5	59.5	59.8	59.9
	Ln	99.4	99.7	99.8	99.8	99.8	99.9	99.9	99.9	99.9
	t	22.9	24.2	24.6	24.9	25.1	25.3	25.4	25.6	25.7
D	B	—	6.7	—	7.0	—	6.9	—	7.1	—
	C	—	91.7	—	92.6	—	93.0	—	93.2	—
	Γ	—	29.4	—	30.7	—	32.2	—	32.3	—
	Ln	—	95.6	—	96.5	—	97.0	—	97.2	—
	t	—	20.3	—	21.3	—	21.9	—	22.2	—
FB	B	4.7	4.5	4.4	4.4	4.3	4.2	4.2	4.2	4.3
	C	92.5	93.4	93.6	93.7	93.8	93.8	93.9	93.8	93.9
	Γ	52.2	55.4	56.4	57.1	57.7	58.4	58.5	58.8	58.9
	Ln	99.3	99.7	99.8	99.8	99.8	99.9	99.8	99.9	99.9
	t	23.3	24.7	25.0	25.1	25.4	25.6	25.9	26.0	26.1

(cont.)

Table 5. Continued

		Location-scale model, $n = 25$								
MC reps.		19	39	59	79	99	199	299	399	499
JB	B	3.1	2.5	2.2	2.2	2.1	1.9	1.9	2.0	2.0
	C	89.2	90.8	91.2	91.4	91.4	91.6	91.7	91.7	91.7
	Γ	43.2	45.6	46.6	47.4	47.8	48.5	48.6	48.6	48.8
	Ln	94.3	96.3	97.1	97.3	97.5	97.9	97.9	98.0	98.0
	t	23.8	25.1	25.8	26.1	26.3	26.6	26.5	26.8	26.7
JB_k	B	4.5	3.8	3.2	3.1	3.0	2.6	2.6	2.6	2.5
	C	88.8	90.3	90.7	90.8	91.0	91.3	91.3	91.3	91.4
	Γ	45.5	48.8	50.0	50.6	51.2	52.0	52.2	52.4	52.4
	Ln	94.9	97.0	97.7	98.0	98.2	98.7	98.8	98.8	99.0
	t	22.8	24.2	25.0	25.4	25.6	26.0	26.0	26.2	26.1

(cont.)

WB and FB are biased in the case of the beta distribution. Although the D test typically shows less power than the other tests in its class, it is not biased in small samples, unlike the SF-type counterparts. The AD outperforms all EDF statistics, compares favorably to the moment tests and has no bias problems. While it is biased against the beta distribution for $n = 25$, the JB almost achieves maximum power against the Cauchy, lognormal and Student *t(5)* distribution; it is outperformed by the AD statistic in the case of the Γ distribution when $n \leq 50$. As expected, all MC tests have very good power when the errors follow the Cauchy and the lognormal distribution even in small samples. Finally, from Table 5, we observe that the number of replications beyond 99 has no significant effect on the power of MC tests.

The regression model. From the results in Table 4, it can be seen that the performance of the regression-based tests can be greatly affected by the design matrix especially for samples of size less than 100. However, it appears that the design matrix has little effect on the ranking of the tests. Furthermore, the results on relative power across tests seem to agree with our findings regarding the location-scale model. In general, the SW-type criteria appear to be the best available; the D statistic is on the whole less powerful than these but is consistently unbiased. The most powerful EDF statistic is the AD; it performs well in comparison with the correlation statistics except perhaps in the $\Gamma(2, 1)$ case. The JB-type tests based on either s or $\hat{\sigma}$ compare favorably to the correlation tests. However, there is no clear indication as to which estimate of σ should be used in practice. The MLE-based JB criterion performs better against the Cauchy, lognormal and the t (5) alternatives, while JB_k appears better for other distributions and is consistently unbiased. For the beta(2,3) alternative, the JB criterion is severely biased for all samples sizes, yet JB_k performs best in comparison with all tests.

5. CONCLUSION

In this paper, we have proposed simulation-based procedures to derive exact *p*-values for several well-known normality tests in linear regression models. Most conventional test procedures were derived in location-scale contexts yet remain asymptotically valid when computed from regression residuals. Here, we have exploited the fact that standard test criteria are pivotal under the null, which allows one to apply the technique of MC tests. The feasibility of the approach suggested was illustrated through a simulation experiment. The results show that asymptotic normality tests are indeed highly unreliable; in contrast MC tests achieve perfect size control and have good power. It is important to emphasize that MC test procedures are not, with modern computer facilities, computationally expensive.

The above findings mean that tables of critical points are no longer required to implement normality tests. Much of the theoretical work in this context has focused on deriving these tables; the reason is clearly the intractable nature of the relevant null distributions. Here we showed that the technique of MC tests easily solves this problem and yields much more reliable procedures.

ACKNOWLEDGEMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and by the Government of Québec (Fonds FCAR). We thank Serge Tardif, Judith Giles and Victoria Zinde-Walsh, the Editor Grayham Mizon and two anonymous referees for several useful comments.

REFERENCES

Affleck-Graves, J. and B. McDonald (1989). Nonnormalities and tests of asset pricing theories. *J. Finance 44*, 889–908.

Anderson, G. (1994). Simple tests of distributional form. *J. Econometrics 62*, 265–276.

Anderson, T. W. and D. A. Darling (1954). A test of goodness-of-fit. *J. Am. Stat. Assoc. 49*, 765–769.

Andrews, D. (1994). Empirical process methods in econometrics. In R. F. Engle and D. L. McFadden (Eds), *Handbook of Econometrics, Volume IV*. Amsterdam: Elsevier Science, pp. 2247–2294.

Andrews, D. W. K. (1988a). Chi-square diagnostic tests for econometric models: introduction and applications. *J. Econometrics 37*, 135–156.

Andrews, D. W. K. (1988b). Chi-square diagnostic tests for econometric models: theory. *Econometrica 56*, 1419–1453.

Andrews, D. W. K. (1997). A conditional Kolmogorov test. *Econometrica 65*, 1097–1128.

Baringhaus, L., R. Danschke, and N. Henze (1989). Recent and classical tests for normality — a comparative study. *Commun. Stat., Simulation Comput. 18*, 363–379.

Barnard, G. A. (1963). Comment on ‘The Spectral Analysis of Point Processes’ by M. S. Bartlett. *J. Roy. Stat. Soc., B 25*, 294.

Bera, A. K., C. M. Jarque, and L.-F. Lee (1984). Testing the normality assumption in limited dependent variable models. *Int. Economic Rev. 25*, 563–578.

Beran, R. and P. W. Millar (1989). A stochastic minimum distance test for multivariate parametric models. *Ann. Stat. 17*, 125–140.

Birnbaum, Z. W. (1974). Computers and unconventional test-statistics. In F. Proschan and R. J. Serfling (Eds), *Reliability and Biometry*. Philadelphia: SIAM, pp. 441–458.

Bowman, K. O. and B. R. Shenton (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika 52*, 591–611.

Cramér, H. (1928). On the composition of elementary errors. *Skandinavisk Aktuarietidskrift 11*, 141–180.

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

D'Agostino, R. B. (1972). Small sample points for the D test of normality. *Biometrika 59*, 219–221.

D'Agostino, R. B. and M. I. A. Stephens (Eds). (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

Dufour, J.-M. (1995). *Monte Carlo tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics*, Technical report, C.R.D.E., Université de Montréal.

Dufour, J.-M., A. Farhat, L. Gardiol, and L. Khalaf (1997). Simulation-Based Finite Sample Normality Tests in Linear Regressions, Technical report, C.R.D.E., Université de Montréal.

Dufour, J.-M. and J. F. Kiviet (1996). Exact tests for structural change in first-order dynamic models. *J. Econometrics 70*, 39–68.

Dufour, J.-M. and J. F. Kiviet (1998). Exact inference methods for first-order autoregressive distributed lag models. *Econometrica 66*, 79–104.

Durbin, J. (1973a): *Distribution Theory for Tests Based on the Sample Distribution Function*. Philadelphia, PA: SIAM.

Durbin, J. (1973b). Weak convergence for the sample distribution function when parameters are estimated. *Ann. Stat. 1*, 279–290.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Ann. Math. Stat. 28*, 181–187.

Efron, B. (1982): *The Jacknife, the Bootstrap and Other Resampling Plans*, CBS-NSF Regional Conference Series in Applied Mathematics, Monograph No. 38. Philadelphia, PA: SIAM.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. New York: Chapman & Hall.

Fama, E. (1976). *Foundations of Finance*. New York: Basic Books.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics 17*, 111–117.

Hall, A. (1990). Lagrange multiplier tests for normality against seminonparametric alternatives. *J. Business and Economic Statistics 8*, 417–426.

Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Harris, L. E. (1986). Cross-security tests of the mixture of distributions hypothesis, *J. Financial Quant. Anal.*, 21, 39–46.

Henze, N. (1996). Empirical-distribution-function goodness-of-fit tests for discrete models. *Can. J. Stat. 24*, 81–93.

Huang, C. J. and B. W. Bolch (1974). On the testing of regression disturbances for normality, *J. Am. Stat. Ass. 69*, 330–335.

Jarque, C. M. and A. K. Bera (1980). Efficiency tests for normality, heteroscedasticity and serial independence of regression residuals, *Econ. Lett. 6*, 255–259.

Jarque, C. M. and A. K. Bera (1987). A test for normality of observations and regression residuals. *Int. Stat. Rev. 55*, 163–172.

Jeong, J. and G. S. Maddala (1993). A Perspective on application of bootstrap methods in econometrics. In G. S. Maddala, C. R. Rao, and H. D. Vinod (Eds), *Handbook of Statistics, Volume 11, Econometrics*. Amsterdam: North Holland, pp. 573–610.

Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Ann. Stat. 14*, 336–347.

Kiviet, J. and J.-M. Dufour (1997). Exact tests in single equation autoregressive distributed lag models. *J. Econometrics 80*, 325–353.

Kolmogorov, A. N. (1933). Sulla determinazione empiricadi una legge di distribuzione. *Giorna. Ist. Attuari. 4*, 83–91.

Lee, L.-F. (1982). Test for normality in the econometric disequilibrium market model. *J. Econometrics 19*, 109–123.

Lilliefors, H. W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown, *J. Am. Stat. Assoc. 62*, 399–402.

Linton, O. and P. Gozalo (1997). *Conditional Independence Restrictions: Testing and Estimation*, Technical report, Cowles Foundation for Research in Economics, Yale University.

Loynes, R. M. (1980). The empirical distribution function of residuals from generalized regression. *Ann. Stat. 8*, 285–298.

Mammen, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Stat. 24*, 307–335.

Mardia, K. V. (1980). Tests of univariate and multivariate normality. In by P. R. Krishnaiah (Eds), *Handbook of Statistics, Volume 1: Analysis of Variance*, Amsterdam: North Holland, pp. 279–320.

Meester, S. G. and R. A. Lockhart (1988). Testing for normal errors in designs with many blocks. *Biometrika 75*, 569–575.

Mukantseva, L. A. (1977). Testing normality in one-dimensional and multi-dimensional linear regression. *Theory of Probability and its Applications 22*, 591–602.

Pfaffenberger, R. C. and T. E. Dielman (1991). Testing normality of regression disturbances: a Monte Carlo study of the Filliben tests. *Comput. Stat. 86*, 359–373.

Pierce, D. A. and R. J. Gray (1982). Testing normality of errors in regression models. *Biometrika 69*, 233–236.

Pierce, D. A. and K. J. Kopecky (1979). Testing goodness-of-fit for the distribution of errors in regression models. *Biometrika 66*, 1–5.

Poirier, D., D. Tello, and S. E. Zin (1986). A diagnostic test for normality within the power exponential family. *J. Business and Economic Statistics 86*, 359–373.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer-Verlag.

Richardson, M. and T. Smith (1993). A test for multivariate normality in stock returns. *J. Business 66*, 295–321.

Royston, J. P. (1982a). Algorithm AS 177. Expected normal order statistics (exact and approximate). *Appl. Stat. 31*, 161–165.

Royston, J. P. (1982b). Algorithm AS 181. The *W* test for normality. *Appl. Stat. 31*, 176– 180.

Royston, J. P. (1982c). An extension of Shapiro Wilks's test for normality to large samples. *Appl. Stat. 31*, 115–124.

Shao, S. and D. Tu (1995): *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Shapiro, S. S. and R. S. Francia (1972). An approximate analysis of variance test for normality. *J. Am. Stat. Assoc. 67*, 215–216.

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 52*, 591–611.

Smirnov, N. V. (1939). Sur les écarts de la courbe de distribution empirique (Russian/French Summary). *Matematičeskiĭ Sbornik N. S. 6*, 3–26.

Stephens, M. A. (1974). EDF Statistics for goodness-of-fit and some comparisons. *J. Am. Stat. Assoc. 69*, 730–737.

Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Ann. Stat. 4*, 357–369.

Stute, W., W. G. Manteiga, and M. N. Quindimil (1993). Bootstrap based goodness-of-fit tests. *Metrika 40*, 243–256.

Vinod, H. D. (1993). Bootstrap methods: applications in econometrics. In G. S. Maddala, C. R. Rao, and H. D. Vinod (Eds), *Handbook of Statistics, Volume 11, Econometrics*, Amsterdam: North Holland, pp. 629–661.

Weisberg, S. (1980). Comment on: some large sample tests for non-normality in the linear regression model, by H. White and G.M. MacDonald, *J. Am. Stat. Assoc. 75*, 28–31.

Weisberg, S. and C. Bingham (1975). An approximate analysis of variance test for non-normality suitable for machine calculation. *Technometrics 17*, 133–134.

White, H. and G. M. MacDonald (1980). Some large sample tests for non-normality in the linear regression model. *J. Am. Stat. Assoc. 75*, 16–28.