Metodološki zvezki, Vol. 16, No. 1, 2019, 1–16

An Empirical Likelihood Ratio Based Comparative Study on Tests for Normality of Residuals in Linear Models

Chioneso Show Marange[1] Yongsong Qin[2]

Abstract

The application of goodness-of-fit (GoF) tests in linear regression modeling is a common practice in applied statistical sciences. For instance, in simple linear regression the assumption of normality of residuals is always necessary to test before making any further inferences. The growing popularity of the use of powerful and efficient empirical likelihood ratio (ELR) based GoF tests in checking for departures from normality in various continuous distributions can be of great use in checking for distributional assumptions of residuals in linear models. Motivated by the attractive properties of the ELR based GoF tests the researchers conducted an extensive Type I error rate assessment as well as a Monte Carlo power comparison of selected ELR GoF tests with well-known existing tests against symmetric and asymmetric alternative OLS and BLUS residuals. Under the simulated scenarios, all the studied tests have good control of Type I error rates. The Monte Carlo experiments revealed the superiority of the ELR GoF tests under certain alternatives of both the OLS and BLUS residuals. Our findings also demonstrated the superiority of OLS over BLUS residuals when one is testing for normality in simple linear regression models. A real data study further revealed the applicability of the ELR based GoF tests in testing normality of residuals in linear regression models.

1 Introduction

The importance of distributional assumptions, especially normality is crucial since it is a fundamental assumption in residual analysis for linear regression models. When such distributional assumptions are not fulfilled, then the inferences and interpretation may not be reliable or valid. In testing for normality, the most commonly used goodness-of-fit (GoF) tests includes the Shapiro-Wilk (SW) test (Shapiro and Wilk, 1965), the modified Kolmogorov-Smirnov (LL) test (Lilliefors, 1967), the Anderson and Darling (AD) test (Anderson and Darling, 1952, 1954) and the Cramér-von Mises (CVM) test (Cramér, 1928; von Mises, 1931 and Smirnov, 1936). The Shapiro-Wilk test has been

[1]Department of Statistics and Biostatistics, Faculty of Science and Agriculture, Fort Hare University, East London, South Africa; cmarange@ufh.ac.za

[2]Department of Statistics and Biostatistics, Faculty of Science and Agriculture, Fort Hare University, Alice, South Africa; yqin@ufh.ac.za

found to outperform other tests (e.g., Razali and Wah, 2011). However, of recent, new GoF tests for normality that utilize the empirical likelihood ratio (ELR) technique (Owen, 2001) are beginning to gain popularity. These test are known to be powerful and efficient tests for normality (e.g., Dong and Giles, 2007; Vexler and Gurevich, 2010; Shan et al., 2010). These tests have proved to outperform other classical established tests, including the Shapiro-Wilk test under certain alternatives.

However, these ELR tests have not yet been applied in testing for normality of residuals in linear regression models. Let us consider a classical linear regression model in its matrix form given by

$$Y = X\beta + \varepsilon, \tag{1.1}$$

where Y is an $n \times 1$ vector of response variables and X is a known $n \times k$ non-stochastic matrix of rank k. The vector β is a $k \times 1$ of unknown regression coefficients whilst ε is an $n \times 1$ vector of unobservable elements. In practice, especially in simple linear regression modeling, the assumption for normality of the error terms is always necessary to check before any further inferences can be done. There are several ways of checking this distributional assumption but in this study we focused on a numerical assessment using GoF tests where the null and alternative hypothesis are given by

H_0 : The errors follow a normal distribution.

H_1 : The errors do not follow a normal distribution.

Since ε is unobservable, GoF tests for ε in linear regression models usually depend on sample errors such as the ordinary least squares residuals (OLS) or the best linear unbiased scalar residuals (BLUS) among others. Most goodness-of-fit tests assume that elements are independent and identically distributed. This proves not to be the case for the OLS residuals in the univariate linear model because these residuals are not independent. The OLS residual vector from a linear regression model is defined as a linear transformation of the response vector Y and can also be expressed as a linear transformation of the error vector ε:

$$\hat{\varepsilon} := MY = M\varepsilon,$$

where $M = I - X(X'X)^{-1}X'$ is an $n \times n$ idempotent symmetric matrix with a rank of $(n - k)$, which annihilates the image of X and preserves its orthogonal complement. Observe that $\mathrm{E}(\hat{\varepsilon}) = 0$ and $\mathrm{Var}(\hat{\varepsilon}) = \sigma^2 M$. Moreover, if ε is normal, so is $\hat{\varepsilon}$. The covariance matrix of the OLS residual vector is not a diagonal matrix but a singular and hence, the elements of the OLS residual vector are not independently distributed. Due to this shortfall of OLS residuals, Theil (1965, 1968) formulated the best linear, unbiased, scalar-type (BLUS) variance residuals for linear regression models. Like the OLS, the BLUS residual vector is defined as a linear transformation of the response vector Y and can also be expressed as a linear transformation of the error vector ε:

$$\varepsilon^* := AY = A\varepsilon,$$

where A is an $(n-k) \times n$ matrix, which, like M, annihilates the image of X, but, in contrast to M, maps its orthogonal complement isometrically onto $\mathbb{R}^{n-k}$. Like for the OLS, we have $\mathrm{E}(\varepsilon^*) = 0$, but in contrast, the covariance matrix of the BLUS residual vector is of full rank and diagonal: $\mathrm{Var}(\varepsilon^*) = \sigma^2 I_{n-k}$. It is normally distributed $N(0, \sigma^2 I_{n-k})$

if and only if the error terms are from a normal distribution and this makes the BLUS residuals ideal for conducting GoF testing (Huang and Bolch, 1974). Despite their desirable theoretical properties, the BLUS residuals are not much used by researchers, perhaps because of computational difficulties. However, when the error terms are not normal, the BLUS residuals may suffer from lack of independence and this may be at least as equal as the lack of independence among OLS residuals.

Standard tests for normality are appropriate for independent data; hence the issue of dependency of these residuals then raises an important question as to which of the tests is most powerful to utilize under the presence of correlations amongst these residuals. Huang and Bolch (1974) conducted a study to compare some well-known GoF tests in testing normality of ordinary least square (OLS) and best linear unbiased scalar (BLUS) residuals in linear models. Their findings revealed that the Shapiro-Wilk test is by and large better than other tests considered and this is in concurrence with Shapiro et al. (1965). The researchers also revealed that the OLS residuals dominated in power as compared to the BLUS residuals. Their findings are similar to those of Ramsey (1969, 1972, 1974).

We conducted an extensive comparison on the performance of the recently proposed ELR based tests to that of other classical well-known existing tests in normality testing of OLS and BLUS residuals in simple linear regression models. Thus, the study investigated the power and empirical probability of Type I error of the selected tests. The study focused on six tests, that is, the modified Kolmogorov-Smirnov (known as the Lilliefors (LL) test) (Lilliefors, 1967), the Anderson and Darling (AD) test (Anderson and Darling, 1952, 1954), the Cramér-von Mises (CVM) test (Cramér, 1928; von Mises, 1931 and Smirnov, 1936), the Shapiro-Wilk (SW) test (Shapiro and Wilk, 1965), the density based empirical likelihood ratio test (Vexler and Gurevich, 2010) and the moment based empirical likelihood ratio based GoF test (Shan et al., 2010). Monte-Carlo simulations using the R statistical package revealed that the ELR tests are superior under certain alternatives of both the OLS and BLUS residuals. A real data study was also utilized.

2 Tests for Normality

Pearson (1895) pioneered the development of methods to test for departures from normality and to date there are numerous GoF tests readily available. For a detailed overview of these tests one can refer to Thode (2002). Several authors have done some investigations and comparisons on the performance of these tests in terms of the power and the probability of Type I error (see for example Shapiro et al., 1968; Huang and Bolch, 1974; Pearson et al., 1977; Dufour et al., 1998; Thode, 2002; Yazici and Yolacan, 2007; Razali and Wah, 2011; Yap and Sim, 2011). Most of these studies have reported that the Shapiro-Wilk test is considered as the better alternative in testing for normality both for continuous data and in residual analysis. This section will present a brief synopsis of the tests considered in this study including the recently proposed ELR based tests for normality that have not yet been applied in residual analysis. The choice of the well-known existing tests was based on a selection of the most efficient and powerful tests that are commonly used by researchers in testing for normality. It should be noted that all tests considered assumes that sample observations are independent and identically distributed.

2.1 Empirical Distribution Function (EDF) Tests

The concept of the EDF tests in assessing for departures from normality in goodness-of-fit testing is focused on comparing the EDF (computed using the observations) with the cumulative distribution function (CDF) of the normal distribution to determine whether there exists a close match between the two functions. In this study we focused on the common EDF tests which are, the modified Kolmogorov-Smirnov (denoted by LL) test (Kolmogorov, 1933; Lilliefors, 1967), the Anderson and Darling (AD) test (Anderson and Darling, 1954) and the Cramér-von Mises (CVM) test (Cramér, 1928; von Mises, 1931 and Smirnov, 1936).

2.1.1 The Modified Kolmogorov-Smirnov Test

The Lilliefors (LL) test is known to be related to the Kolmogorov-Smirnov (KS) test where it is regarded as a modified version of the KS test. Developed by Lilliefors (Lilliefors, 1967), this test compares the EDF of the sample observations with a normal distribution where its unknown mean and standard deviation are first estimated from the data. The major difference between the Lilliefors (LL) and Kolmogorov-Smirnov (KS) test statistic is that the EDF from the LL test is obtained from standardized sample observations while the KS test uses the observed values. The LL statistic is defined as

$$LL = \sup_{x \in \mathbb{R}} |F_n(x) - F^*(x)|,$$

where $F_n(x)$ is the empirical CDF whilst $F^*(x)$ is the hypothesized CDF. The LL test is readily available in several statistical packages. In this study we used the function `lillie.test()` which is available in the nortest R statistical package.

2.1.2 Cramér-von Mises (CVM) Test

The Cramér-von Mises (CVM) test is one of the well-known EDF tests developed by Cramér (1928), von Mises (1931) and Smirnov (1936). The CVM test statistic is distribution free, that is, the distribution is independent of the hypothesized distribution function, $F^*(x)$. The CVM test statistic can be given by

$$CVM = n \int_{-\infty}^{\infty} [F_n(x) - F^*(x)]^2 dF^*(x).$$

where $F_n(x)$ is the empirical CDF. The CVM test rejects H_0 if $CVM \geq C_{1-\alpha}$, where the critical values $(C_{1-\alpha})$ are easily obtained (one can check in Anderson and Darling, 1954). The `cvm.test()` in the nortest R statistical package was used to implement the CVM GoF test.

2.1.3 Anderson and Darling (AD) Test

The Anderson and Darling (AD) test is a modified version of the Cramér-von Mises (CVM) test and is considered to be the most powerful EDF test (Arshad et al., 2003). The difference between the AD and CVM test is based entirely on the fact that the AD

test statistic is more sensitive and focuses more heavily on the weight of the normal distribution tails (Farrel and Stewart, 2006) like in the CVM test smaller values indicate that the distribution is consistent with a normal distribution. One major drawback of the AD test is on the calculation of the critical values which are required to be computed for each specified distribution. Anderson and Darling (1954) defined the test statistic as

$$A^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F^*(x)]^2}{F^*(x)(1 - F^*(x))} dF^*(x),$$

where $F_n(x)$ is the empirical CDF and $F^*(x)$ is the cumulative distribution function of the null distribution. This is a weighted average of the squared difference $[F_n(x) - F^*(x)]^2$, which is weighted by $\psi(x)$. The weight function $\psi(x)$ is non-negative which is computed by $\psi(x) = [F^*(x)(1 - F^*(x))]^{-1}$. It should be noted that when $\psi(x) = 1$, the AD test statistic becomes the CVM test statistic. In order to reject the null hypothesis at a specified level of significance (α), the test statistic, A^2, should be greater than the critical value that is obtained from Monte Carlo simulations. The `ad.test()` which is available in the nortest R statistical package was used to implement the AD GoF test.

2.2 Regression and Correlation Tests

Another category of normality tests that was considered in this study is the regression and correlation tests. These tests are entirely based on the ratio of two weighted least squares estimates of scale obtained from order statistics. This study only focused on regression tests. The most common regression test is the one developed by Shapiro and Wilk (1965).

2.2.1 The Shapiro-Wilk (SW) Test

The Shapiro-Wilk (SW) test was developed by Shapiro and Wilk (1965) and is regarded by most researchers as the best choice for normality testing (e.g., Thode, 2002). It has become the preferred GoF test for normality in residual analysis for linear regression models and other statistical applications due to its desirable power properties (Mendes and Pala, 2003). Given an ordered sample of n sample observations, that is, $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ the SW test proposed by Shapiro and Wilk (1965) uses the test statistic

$$SW = \frac{(\sum_{i=1}^{n} a_i x^{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.1}$$

where x_i is the i^{th} order statistic, $\bar{x}$ is the mean of the sample observations and a_i values are computed using the sample observation's (x_i) means, variances and covariances. Thus

$$a_i = (a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}},$$

where V is the covariance matrix and m^T are the expected values of the order statistics of $i.i.d.$ sample observations from a standard normal distribution. The value of the test statistic is between 0 and 1, where small values will result to H_0 being rejected. The test was originally restricted for sample size of less than 50. Since then, extensive research

has been done to modify the SW test. Royston (1982) modified the SW test in order to widen the constraint of the sample size to 2000. He further observed that the SW test had weaknesses on the approximation of weights that are utilized in the algorithms for sample sizes greater than 50. Royston (1995) then proposed an improved modification of the approximation to the weights which can cater for any sample size in the range $3 \leq n \leq 5000$. The Shapiro-Wilk test is available in several statistical packages. This study used the stats R statistical package utilizing the function `sw.test()`.

2.3 Empirical Likelihood Ratio (ELR) Based Tests

The ELR based GoF tests have recently gained popularity and are based upon the empirical likelihood function (DiCiccio et al., 1989; Owen, 1988, 1991, 2001; Dong and Giles, 2007; Vexler and Gurevich, 2010; Shan et al., 2010; Yu et al., 2010). Recently, several GoF tests for normality have been proposed using the empirical likelihood methodology. In this study we focused on a classical ELR GoF test based on moment constraints (proposed by Shan et al., 2010) as well as a density based ELR test (proposed by Vexler and Gurevich, 2010). These tests are known to be efficient and powerful with critical values that can be easily computed using Monte-Carlo simulations.

2.3.1 Classical Empirical Likelihood Ratio Based Test

Under this category we are going to focus on a recently developed test by Shan et al. (2010) to test for departures from normality based on moment relations of a standard normal distribution. Shan et al. (2010) proposed this method after identifying the weaknesses in a method that was developed by Dong and Giles (2007). Dong and Giles (2007) proposed an empirical likelihood GoF test statistic for normality by using the method presented by Owen (2001). They used the first four moment constraints (that is, the mean, variance, skewness and kurtosis) of the normal distribution. However, due to the fact that the test involves numerically complex nonlinear equations it is not easy to utilize. Also, the numerical convergence of the global maximum is not certain. In addition, Shan et al. (2010) also noted that the asymptotic Type I error rate for the classical ELR test by Dong and Giles (2007) has poor control in small samples. Shan et al. (2010) then developed a simple and efficient ELR GoF test (SEELR) for normality which is rooted in the dependence of the moment constraints that are related to the standard normal distribution.

To summarize this test, consider n unordered independent and identically distributed sample observations, i.e., $X_1, X_2, \ldots, X_n$. The SEELR tests the null hypothesis that the data follow a normal distribution with mean μ and variance σ^2. In this case, both μ and σ are unknown fixed parameters. The test makes use of standardized sample observations using the Lin and Mudholkar (1980) transformation. By definition, the standardized random variables $Z_1, Z_2, \ldots, Z_n$ follow a t-distribution with $n-2$ degrees of freedom. It can be easily noted that as the degrees of freedom become large, the t-distribution approaches a standard normal distribution and the standardized sample observations become asymptotically independent. Shan et al. (2010) then proposed to use the moment function of the t-distribution with $n-2$ degrees of freedom. Using the empirical likelihood tests under the null hypothesis:

$$H_0 : \mathrm{E}(Z^k) = E(T^k_{n-2}),$$

where T^k_{n-2} follows a t-distribution with $n-2$ degrees of freedom. Following the utilization of the EL methodology, the researchers proposed to reject the null hypothesis if

$$\text{SEELR} := \max_{k \in G}(-2LLR)_k > C_\alpha,$$

where C_α is the test threshold, and G is a set of integer values. For high levels of power under the null hypothesis G was set to $\{3,4,5,7\}$. For more details on the SEELR test one can refer to Shan et al. (2010). In this study we utilized this test using the R statistical package. The R-code is available in the author's article.

2.3.2 Density Based Empirical Likelihood Based Test

The density based empirical likelihood ratio test (*dbEmpLikeGoF*) is a relatively recent technique which has significantly outperformed most classical existing methods (Vexler and Gurevich, 2010). The *dbEmpLikeGoF* technique was successfully applied to develop powerful and efficient GoF tests for one and two-sample problems (Vexler and Gurevich, 2010; Vexler et al., 2011; Karagrigoriou, 2012; Gurevich and Vexler, 2011; Vexler et al., 2012). These tests offer a variety of GoF tests for distributional assumptions from a wide range of hypotheses. In this study we focused on the density-based EL ratio test for normality and the derivations can be found in Vexler and Gurevich (2010). Vexler and Gurevich (2010) considered the following EL ratio test statistic

$$V_n = \min_{1 \le m < n^{1-\delta}} (2\pi s^2)^{n/2} \prod_{i=1}^{n} \frac{2m}{n(X_{(i+m)} - X_{(i-m)})},\ 0 < \delta < 1,$$

where s is the sample standard deviation and $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ are the order statistics of the sample $X_1, X_2, \ldots, X_n$. The null hypothesis is rejected if and only if

$$\log(V_n) > C,$$

where C is the test threshold and V_n is the test statistic defined above. Miecznikowski et al. (2013) presented an R package for statistical tests that are based on the *dbEmpLikeGoF* technique. This is the package that was utilized in this study.

3 Monte Carlo Simulation Procedures

This section outlines the Monte Carlo simulation procedures that were considered for GoF power comparisons in testing for normality of OLS ($\hat{\varepsilon}$) and BLUS residuals (ε^*) in a univariate linear regression model with the form presented earlier in (1.1). Previous studies have considered evaluating multiple linear regression models whereby there is a constant term plus at least two or more regressors (for example see, Huang and Bolch, 1974; Ramsey, 1974; Weisberg, 1980; White and MacDonald, 1980; Jarque and Bera, 1987). Jarque and Bera (1987) considered regressors $X_1, \ldots, X_4$ following the study of White and MacDonald (1980). For their experiments they decided to set $X_{1i} = 1$ $(i = 1, 2, \ldots, n)$ and then generate X_2, X_3 and X_4 from a uniform distribution. On the other hand, Huang and Bolch (1974) also considered a multiple linear regression model with a

constant term and three additional regressors that were drawn from a uniform distribution and held constant for each experiment. However, in their experiments Huang and Bolch (1974) proposed to use the following pre-defined model

$$Y_i = -20.0 + 4.5X_{1i} - 1.5X_{2i} + 2.8X_{3i} + \varepsilon_i,\ i = 1, 2, \ldots, n.$$

Due to the different forms in which the regressors can be generated and/or due to changes in the number of regressors k, Weisberg (1980) as well as Jarque and Bera (1987) found that the power of tests may vary. However, the power ranking of the tests does not change (Weisberg, 1980; Jarque and Bera, 1987). Therefore, we then proposed not to investigate the effect of the number of regressors k, and the elements of the regressor matrix but rather focused our attention more on the distribution of the residual vector and sample size. Following the approach by Huang and Bolch (1974) we then proposed to use a pre-defined simple linear regression model of the form

$$Y_i = 1 + 2X_{1i} + \varepsilon_i,\ i = 1, 2, ..., n.$$

The regressor, which is independent of ε_i was randomly generated once from a uniform distribution and kept constant for each simulation based upon a given sample size. For the assessment of Type I error, the random disturbances were drawn from a standard normal distribution. In terms of power simulations, the resulting vector ε has a specified known distribution. Thus, in computing the power of a test, the error vectors of the random disturbances (ε) were drawn from four alternative distributions, Exponential (1), Lognormal (0,1), Cauchy (0,1) and Uniform (0,1) distributed OLS and BLUS residuals. The symmetric and asymmetric nature of these distributions as well as their different shapes offer a variety of contrasts to the normal distribution.

Monte Carlo procedures were then used to evaluate the probability of the Type I error and the power of the Lilliefors (LL) test (Lilliefors, 1967), the Anderson and Darling (AD) test (Anderson and Darling, 1952, 1954), the Cramér-von Mises (CVM) test (Cramér, 1928; von Mises, 1931 and Smirnov, 1936), the Shapiro-Wilk (SW) test (Shapiro and Wilk, 1965), the density based empirical likelihood ratio based test (DB) (Vexler and Gurevich, 2010) and the simple and exact empirical likelihood ratio test based on moment relations (SEELR) (Shan et al., 2010). These GoF tests are all directly applicable to the OLS and BLUS residuals. We used the function `lm()` for the generation of OLS residuals. For the generation of BLUS residuals we used the R Code for Theil's BLUS residuals presented by Vinod (2014). Three levels of significance, α, 1%, 5% and 10% were considered in order to investigate the effect of the level of significance on the power of the tests. Power simulations were conducted using 5000 replications with varying sample sizes (n = 15, 30, 50, 80, 100, 150 and 200), at the various levels of significance.

3.1 Assessing the Probability of the Type I Error of the GoF Tests

Before the power simulation study we assessed the probability of the Type I error of the GoF tests using 500 000 simulations over different α levels (α = 0.01, 0.05 and 0.10) and sample sizes (n = 15, 50 and 150). By definition, a GoF test is intended to reject the null hypothesis with a chance of at most α when the null is true, i.e., false positive rate. We assessed the empirical probabilities of the Type I error for all tests under OLS and

BLUS residuals as well as normally distributed data with zero mean and unit variance. Table 3 presents the simulated probabilities of the Type I error, along with the standard error for all tests. For clarity and comparison sake, Figures 3 to 8 shows the graphical representations of the cumulative Type I error rates only at 0.05 level of significance. The plots for the empirical cumulative probability function of the simulated p-values for $\alpha = 0.01$ and $\alpha = 0.10$ were omitted since their plots were more or less the same as those for $\alpha = 0.05$.

It is clearly evident that the plots produced the expected appearance in most of the simulated scenarios. That is, the plots show close to the α-level of simulated Type I error rates for both the OLS and BLUS residuals as well as the standard normal data. The closeness of the estimated probabilities of Type I error to the nominal value ($\alpha = 0.05$) attests that the GoF test does perform as expected. However, our empirical results from the simulated Type I error rates of the density based ELR test provide evidence which suggests that the test tends to under reject in moderate sample settings (i.e., $n = 50$) at low levels of significance (i.e., $\alpha = 0.01$ and 0.05). This is however of little concern for one to use the test under these settings as the deviation from the true nominal levels is somewhat within a statistically acceptable range. Generally, as expected, the behaviour of the Type I error rates for the BLUS residuals and the standard normal data is the same since both sets of data are known to be independent unlike the OLS residuals which suffer from lack of independence. However, the plots for the OLS residuals are somewhat similar to that of the BLUS residuals and the standard normal data in all tests for $n = 50$ and 150. It is also interesting to note that the estimated Type I error rates for OLS residuals in small sample sizes (i.e., $n = 15$) are generally smaller than those for the BLUS residuals and the standard normal data in our experiments. Also for large sample size, $n = 150$, the DB test tends to give estimated Type I error rates that are consistently higher than the nominal α-levels. However, the ELR based tests are the only tests that have estimated Type I error rates that are consistently closer to the true nominal α-levels for small sample size (i.e., $n = 15$) under both the OLS and BLUS residuals as well as the standard normal data. Generally, from these findings all tests considered in this study can be used to test for normality in OLS and BLUS residuals.

3.2 Power Analysis: Simulation Results

Table 4 gives the results when the alternative distribution is exponential (i.e., Exp (1)). The SEELR test has the highest power among the tests for significance levels of 1%, 5%, 10%. That is, in general, the SEELR test out performed all the tests studied under exponentially distributed OLS and BLUS residuals alternatives. The SW test is the second most powerful test under the exponential alternative. For small sample size (i.e., $n = 15$), at 1% level of significance, the SW test is seen to be superior to the SEELR test. Generally, the power of the DB test is slightly lower to that of the AD test, whilst the LL test has the least power under these exponentially distributed OLS and BLUS residual alternatives. Under Lognormal (0,1) distributed OLS and BLUS residuals (see Table 5) both the SW and the SEELR tests are generally the most superior tests. However, for $\alpha = 0.01$ the AD test is slightly superior than the SEELR test against OLS residuals. It is also important to note that the SEELR tests is the most powerful test under the lognormal (0,1) distributed BLUS residuals for all the different significance levels considered in this study. The power

of the DB test is only superior to that of the LL test under this alternative.

For the symmetric, Cauchy (0,1) distributed OLS and BLUS residuals, the power of the DB and SEELR tests are inferior to that of other tests considered in this study (see Table 6). The AD test is the most powerful, with the SW test being the second most superior but somewhat comparable to the CVM test. For α-levels of 0.05 and 0.10, the powers of all of the AD, SW and CVM tests converge to 100% as n grows, though more slowly for the BLUS residuals. Only under the Cauchy (0,1) alternative is the LL test superior to the ELR based tests. Table 7 gives the results when the alternative distribution is Uniform (0,1). The DB test is the most powerful among all of the six tests considered for all the given α-levels at various sample sizes. For all α-levels, the power of the DB test converges to 100% as n grows, for both the OLS and the BLUS residuals. The SW tests is once again the second most superior test whilst the LL test is the least powerful test under the uniformly distributed OLS and BLUS residuals. The SEELR test is only superior to the CVM and LL tests whilst slightly inferior to the AD test. Generally, when the alternative is symmetric and uniformly distributed, all of the tests have quite low power as compared to other alternatives considered in this study. In summary, as expected, the simulation study shows that none of the tests considered in this study can be considered to be uniformly the best for all the alternative distributions studied (for example see, Janssen, 2000). However, for all the simulated scenarios, the SW test was either the most powerful (i.e., under Lognormal alternative) or the second most powerful (i.e., under Exp (1), Cauchy (0,1) and Uniform (0,1) alternatives). On the other hand, both the ELR tests considered in this study have proved to be the most powerful tests, only under certain alternatives. In terms of the residuals, the OLS outperformed the BLUS residuals in all simulated scenarios.

4 Real Data Example

In order to assess the applicability of the ELR based tests on real data, we conducted a study using the mammal data ($n = 62$) which are data records of average weight of the brain and body for a number of mammal species. This data has been used in several statistical applications in linear regression modelling which includes Spaeth (1991) and Weisberg (1980). In our study we were interested in modelling the effect of brain weight on body weight using a simple linear regression model. The model under consideration can be written as

$$y = \beta_0 + \beta_1 x_1 + \epsilon,$$

where y is body weight, x_1 is the brain weight and $\hat{\epsilon}$ are the residuals. We were interested in testing whether both the OLS and BLUS residuals from this model are consistent with a normal distribution. Figure 1 below shows the plots to assess normality of the OLS and BLUS residuals for the resultant mammal data model.

From the plots it is evident that both the OLS and BLUS residuals are not consistent with the normal distribution. To further check this inconsistency, we carried out a GoF test for normality using the modified KS test, the AD test, the CVM test, the SW test, the DB test and the SEELR test. We took note of the respective p-values of the tests. The results are presented in Table 1 below and it is clear that at 5% level of significance

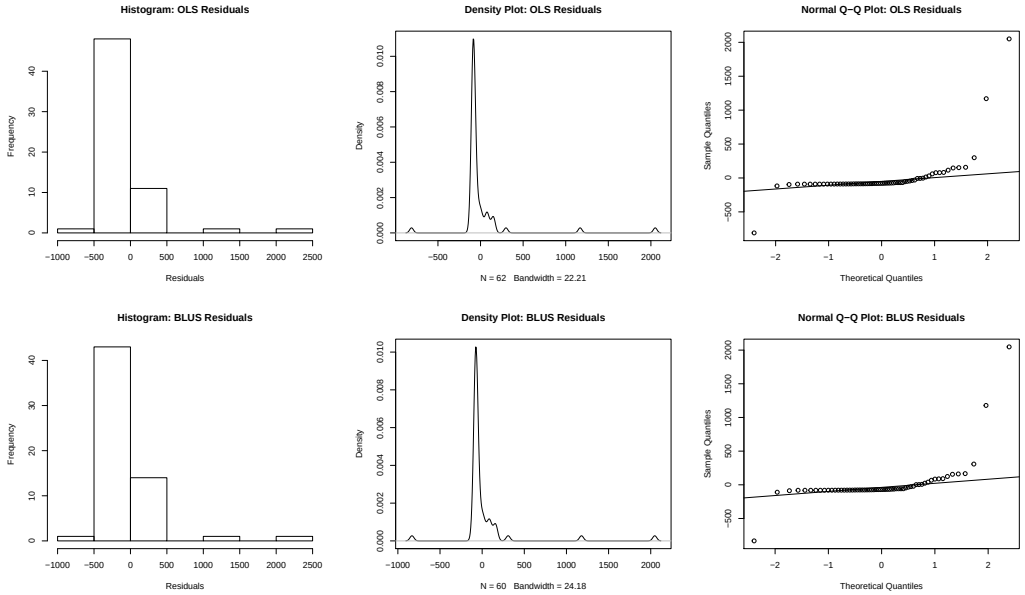

Figure 1: Plots to check for normality on OLS and BLUS residuals of the mammal model

the ELR based tests also rejected the null hypothesis just like any other common existing tests, hence enabling us to conclude that the residuals are not normally distributed.

Table 1: Assessing normality of OLS and BLUS residuals using the mammal data. Presented are p-values for testing normality of residuals ($n = 62$, $\alpha = 0.05$).

Residuals	LL	AD	CVM	SW	DB	SEELR
OLS Residuals	<0.0001	<0.0001	<0.0001	<0.0001	0.0010	<0.0001
BLUS Residuals	<0.0001	<0.0001	<0.0001	<0.0001	0.0010	<0.0001

Note: Testing for normality of OLS and BLUS residuals using, the Lilliefors (LL) test, the Anderson and Darling (AD) test, the Cramér-von Mises (CVM) test, the Shapiro Wilk (SW) test, the Density based empirical likelihood ratio (DB) test and the simple and exact empirical likelihood ratio based (SEELR) test.

In order to normalize the residuals a log transformation of the variables was done. Figure 2 below shows the plots for assessing the OLS and BLUS residuals for the linear model using the log transformed observations. The plots are clearly suggestive that both the OLS and BLUS residuals are from a normal distribution.

Further assessment for normality of these residuals was done by conducting goodness-of-fit tests. Thus, the residuals after the log transformation, should be close to normality, and the tests for normality should provide large p-values. The results in Table 2 below shows that all the tests considered including the ELR based tests suggest that the residuals are now normally distributed as indicated by the plots in Figure 2 above.

This real data example has shown that the ELR based GoF tests are comparable to the studied common existing GoF tests and can be easily applied in real life scenarios.

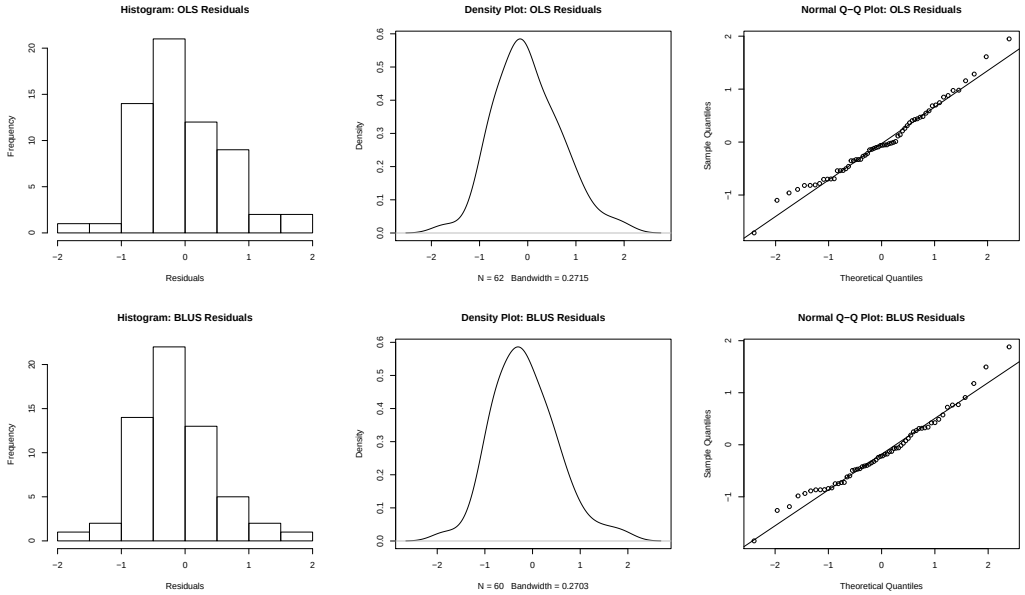

Figure 2: Plots to check for normality on OLS and BLUS residuals of the transformed mammal model.

Table 2: Assessing normality of OLS and BLUS residuals using the log transformed mammal data. Presented are p-values for testing normality of residuals ($n = 62$, $\alpha = 0.05$).

Residuals	LL	AD	CVM	SW	DB	SEELR
OLS Residuals	0.0773	0.3655	0.3095	0.5293	0.9101	0.2582
BLUS Residuals	0.5706	0.3391	0.4040	0.3448	0.9503	0.1382

Note: Testing for normality of OLS and BLUS residuals using, the Lilliefors (LL) test, the Anderson and Darling (AD) test, the Cramér-von Mises (CVM) test, the Shapiro Wilk (SW) test, the Density based empirical likelihood ratio (DB) test and the simple and exact empirical likelihood ratio based (SEELR) test.

5 Conclusion

We have demonstrated the applicability of the ELR based tests in goodness-of-fit testing of normality for residuals in simple linear regression models. The present study confirms previous findings that the Shapiro-Wilk test is overall powerful in GoF testing of residuals in linear regression models (e.g., Shapiro and Wilk, 1965; Dyer, 1974; Huang and Bolch, 1974). However, this study has shown that under certain alternatives, certain ELR based tests outperform the Shapiro-Wilk test. In particular, the SEELR test proposed by Shan et al. (2010) outperforms the Shapiro-Wilk test when the alternative is Exponential (1) whilst the density based ELR test proposed by Vexler and Gurevich (2010) is superior under the Uniform (0,1) distributed OLS and BLUS residuals. Therefore, the ELR based tests seem to be promising alternatives, but they cannot replace the classical tests yet. However, this might be the case after certain improvements. It would be desirable to develop an ELR based test which outperforms the classical tests under most alternative distributions that occur in practice. In particular, further research on the weakness of the

moment based ELR GoF tests against symmetric alternatives needs to be done. It would be interesting, therefore, for future research to explore and implement the techniques that can address this issue and at the same time maintain the good power properties for the ELR approach. We also noticed that the simulated Type I error rates of the density based ELR test provide evidence which suggests that the test tends to under reject in moderate sample settings. This is however of little concern for one to use the test under these settings as the deviation from the true nominal levels is somewhat within a statistically acceptable range.

In terms of the residuals, Huang and Bolch (1974) as well as Ramsey (1974) alluded that OLS residuals are more superior to BLUS residuals when one is testing normality, which is also the case in our study and this finding is also consistent with Ramsey (1969, 1972). The use of transformed residuals, such as the BLUS residuals comes with some computational burdens involved in calculating them. However, since the BLUS residuals may suffer from lack of independence and this may be at least as equal as the lack of independence among OLS residuals when the error terms are not normal, one can just make use of the OLS residuals in testing for normality in simple linear regression models. In other related work, some researchers have rather supported the use of OLS residuals over other forms of transformed residuals (e.g., Jarque and Bera, 1987) whereas some have shown indecisiveness in choosing between OLS and BLUS residuals (e.g., Ramsey, 1969; Ramsey and Gilbert, 1972). However, it will be interesting for future research to look at the applicability of the ELR based tests in GoF testing for normality of other forms of residuals, hence, extensions of our study to more complex linear regression models will be a potential area of future research.

Acknowledgements

We want to thank the Govan Mbeki Research Unit of the hosting university for sponsoring this study. The authors would wish to extend their gratitude to Professor Albert Vexler for his insightful comments on Researchgate. The authors are also thankful to the reviewers and the Editor, whose helpful comments and suggestions contributed to improve the quality of the article.

References

[1] Anderson, T. W. and Darling, D. A. (1952): Asymptotic theory of certain "goodness-of-fit" criteria based stochastic processes. *The Annals of Mathematical Statistics*, **23**(2), 193–212.

[2] Anderson, T. W. and Darling, D. A. (1954): A test of goodness of fit. *Journal of the American Statistical Association*, **49**(268), 765–769.

[3] Arshad, M., Rasool, M.T. and Ahmad, M.I. (2003): Anderson Darling and modified Anderson Darling tests for generalized Pareto distribution. *Pakistan Journal of Applied Sciences*, **3**(2), 85–88.

[4] Cramér, H. (1928): On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, **1928**(1), 13–74.

[5] DiCiccio, T., P. Hall, and J. Romano (1989): Comparison of Parametric and Empirical Likelihood Functions. *Biometrika*, **76**, 465–476.

[6] Dong, L. B. and Giles, D. E. A. (2007): An empirical likelihood ratio test for normality. *Communications in Statistics – Simulation and Computation*, **36**, 197–215.

[7] Dufour, J. M., Farhat, A., Gardiol, L. and Khalaf, L. (1998): Simulation-based finite sample normality tests in linear regressions. *The Econometrics Journal*, **1**(1), 154–173.

[8] Dyer, A. R. (1974): Comparisons of tests for normality with a cautionary note. *Biometrika*, **61**(1), 185–189.

[9] Farrell, P. J. and Rogers-Stewart, K. (2006): Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, **76**(9), 803–816.

[10] Gurevich G. and Vexler A. (2011): A two-sample empirical likelihood ratio test based on samples entropy. *Statistics and Computing*, **21**, 657–670.

[11] Huang, C. J. and Bolch, B. W. (1974): On the testing of regression disturbances for normality. *Journal of the American Statistical Association*, **69**(346), 330–335.

[12] Janssen, A. (2000): Global power functions of goodness of fit tests. *Annals of Statistics*, **28**(1), 239–253.

[13] Jarque, C. M. and Bera, A. K. (1987): A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, **55**(2), 163–172.

[14] Karagrigoriou A. (2012): *Goodness-of-Fit Tests for Reliability Modeling*. New York, NY: Springer.

[15] Kolmogorov, A. N. (1933): Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 83–91.

[16] Lilliefors, H. (1967): On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62**(318), 399–402.

[17] Mendes, M. and Pala, A. (2003): Type I error rate and power of three normality tests. *Pakistan Journal of Information and Technology*, **2**(2), 135–139.

[18] Miecznikowski, J. C., Vexler, A. and Shepherd, L. A. (2013): dbEmpLikeGOF: An R package for nonparametric likelihood-ratio tests for goodness-of-fit and two-sample comparisons based on sample entropy. *Journal of Statistical Software*, **54**(3), 1–19.

[19] Lin, C. C. and Mudholkar, G. S. (1980): A simple test for normality against asymmetric alternatives. *Biometrika*, **67**(2), 455–461.

[20] Owen, A. B. (1988): Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2), 237–249.

[21] Owen, A.B. (1991): Empirical likelihood for linear models. *The Annals of Statistics*, **19**(4), 1725–1747.

[22] Owen, A. B. (2001): *Empirical Likelihood*. New York, NY: Chapman and Hall.

[23] Pearson, K. (1895): Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, **186**, 343–414.

[24] Pearson, E. S., D'Agostino, R. B. and Bowman, K. O. (1977): Tests for departure from normality: Comparison of powers. *Biometrika*, **64**(2), 231–246.

[25] Ramsey, J. B. (1969): Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society*, **31**(2), 350-371.

[26] Ramsey, J. B. (1974): Classical model selection through specification error tests. In P. Zarembka (Ed): *Frontiers in Econometrics*, 13–47. New York, NY: Academic Press.

[27] Ramsey, J. and Gilbert, R. (1972): A Monte Carlo study of some small sample properties of tests for specification error. *Journal of the American Statistical Association*, **67**(337), 180–186.

[28] Razali, N. M. and Wah, Y. B. (2011): Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics*, **2**(1), 21–33.

[29] Royston, J. P. (1982): An extension of Shapiro and Wilk's W test for normality to large samples. *Applied statistics*, **31**(2), 115–124.

[30] Royston, P. (1995): A remark on algorithm AS 181: The W-test for normality. *Journal of the Royal Statistical Society*, **44**(4), 547–551.

[31] Shan, G., Vexler, A., Wilding, G. E. and Hutson, A. D. (2010): Simple and exact empirical likelihood ratio tests for normality based on moment relations. *Communications in Statistics-Simulation and Computation*, **40**(1), 129–146.

[32] Shapiro, S. S. and Wilk, M. B. (1965): An analysis of variance test for normality (complete samples). *Biometrika* **52**(3/4), 591–611.

[33] Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968): A comparative study of various tests for normality. *Journal of the American Statistical Association*, **63**(324), 1343–1372.

[34] Smirnov, N.V. (1936): Sui la distribution de w^2 (Criterium de M.R.v. Mises). *Comptes rendus de l'Académie des Sciences*, **202**, 449–452.

[35] Spaeth H. (1991): *Mathematical Algorithms for Linear Regression*. New Your, NY: Academic Press.

[36] Theil, H. (1965): The analysis of disturbances in regression analysis. *Journal of the American Statistical Association*, **60**(312), 1067–1079.

[37] Theil, H. (1968): A simplification of the BLUS procedure for analyzing regression disturbances. *Journal of the American Statistical Association*, **63**(321), 242–251.

[38] Thode, H.C. (2002): *Testing for normality*. New York, NY: CRC Press.

[39] Vexler, A., Shan, G., Kim, S., Tsai, W. M., Tian, L. and Hutson, A. D. (2011): An empirical likelihood ratio based goodness-of-fit test for inverse Gaussian distributions. *Journal of Statistical Planning and Inference*, **141**(6), 2128–2140.

[40] Vexler, A. and Gurevich, G. (2010): Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Computational Statistics and Data Analysis*, **54**(2), 531–545.

[41] Vexler, A., Tsai, W. M., Gurevich, G., and Yu, J. (2012): Two-sample density-based empirical likelihood ratio tests based on paired data, with application to a treatment study of attention-deficit/hyperactivity disorder and severe mood dysregulation. *Statistics in Medicine*, **31**(17), 1821–1837.

[42] von Mises, R. (1931): *Wahrscheinlichkeitsrechnung und Ihre Anwendung in der Statistik und Theoretischen Physik*. Leipzig: Franz Deuticke.

[43] Vinod, Hrishikesh D. (2014): Theil's BLUS Residuals and R Tools for Testing and Removing Autocorrelation and Heteroscedasticity. Retrieved from `https://ssrn.com/abstract=2412740`.

[44] Weisberg, S. (1980): Comment on paper by H. White and G.M. MacDonald. *Journal of the American Statistical Association*, **75**, 28-31.

[45] White, H. and MacDonald, G. M. (1980): Some large-sample tests for nonnormality in the linear regression model. *Journal of the American Statistical Association*, **75**(369), 16–28.

[46] Yap, B. W. and Sim, C. H. (2011): Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, **81**(12), 2141–2155.

[47] Yazici, B. and Yolacan, S. (2007): A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, **77**(2), 175–183.

[48] Yu, J., Vexler, A. and Tian, L. (2010): Analyzing incomplete data subject to a threshold using empirical likelihood methods: An application to a pneumonia risk study in an ICU setting. *Biometrics*, **66**(1), 123–130.