Journal of Mathematics and Statistics

Original Research Paper

Statistical Size and Power of Eight Normality Tests in Presence of GARCH (1 1) Errors

Julio César Alonso and Daniela Estrada

Department of Economy, Universidad Icesi, Cali, Colombia. Street 18 Num. 122-135 Cali-Colombia

Article history
Received: 02-10-2018
Revised: 19-01-2019
Accepted: 09-02-2019

Corresponding Author:
Julio César Alonso
Department of Economy,
Universidad Icesi, Cali,
Colombia. Street 18 Num. 122-135 Cali-Colombiaa
Email: jcalonso@icesi.edu.co

Abstract: In this work, we assess the power and size of eight normality tests underthe assumption that errors follow a GARCH (1, 1) process by using MonteCarlo simulations. Four results stand out. First, the presence of a GARCH(1, 1) process increases the probability of making type I error. Second, Pearsonnormality test is recommended if it is assumed that errors follow a GARCH(1, 1) process. Third, statistical power varies depending on the type of heteroscedasticity and distribution considered. Fourth, normality tests have lowstatistical power and size (less than or equal to the nominal level) for smalland homoscedastic samples.

Keywords: GARCH, Normality Tests, Statistical Size, Statistical Power, MonteCarlo Simulation

Introduction

Autoregressive conditional heteroscedasticity (ARCH) models that describe heteroscedastic behavior in time series errors were introduced by Engle (1982) more than 36 years ago. Four years later Bollerslev (1986) generalized the ARCH model by introducing the generalized autoregressive conditionally heteroscedastic (GARCH) models.

Nowadays, GARCH models are widely used and, in some contexts, have a better fit than an ARCH model (Enders, 2003). GARCH models have proven very useful for modeling financial time series behavior it solves Ordinary Least Squares (OLS) estimator's inefficiency caused by heteroscedastic errors. This result makes possible to use standard errors, t and F statistics to make inferences (Green, 2012).

It also provides a measure of volatility, on which financial decisions related to risk analysis and portfolio selection are based and can be useful in the analysis of changes in exchange and interest rates (Bollerslev *et al*., 1992).

GARCH models can be estimated by the method of Maximum Likelihood (ML), which assumes that the errors follow a normal distribution. This assumption is essential for some estimation methods, such as ML, but it is also necessary to make an inference from small samples and for constructing prediction intervals of any sample size.

If innovations of GARCH models are expected to follow a distribution different from the normal distribution, the literature suggests a Quasi-Maximum Likelihood (QML) method. However, Engle and Gonzalez-Rivera (1991) showed that estimators lose efficiency if the density function of the error term is not adequately specified.

Other authors have arrived at similar conclusions. Bellini and Bottolo (2008), through Monte Carlo simulations, found that ML and QML estimators underestimated or overestimated volatilities depending on the misspecification assumed. That variability can often generate a spurious "IGARCH effect" when estimating under a weak stationary constraint. Similarly, Klar *et al*. (2012) stated that QML estimators associated with an incorrect specification of the error term might imply a loss of efficiency of the estimators, which could imply a wrong assessment of Value-at-Risk (VaR) and an inaccurate forecast of priced options.

Therefore, the normality assumption is crucial for a practitioner when estimating GARCH models. However, little is known about the statistical power and size of normality formal tests under the presence of errors that do not follow an independent and homoscedastic data generating processes.

As far as the authors are aware, Vavra (2011) and Fiorentini *et al*. (2004) are the only approaches that have studied the performance of normality test under the presence of errors that follow a GARCH process. The first article evaluated three tests of normality (Jarque-Bera (JB), the J test based on the generalized method of moments and a third based on quantiles) and their findings showed that the quantile test has a better performance than JB and that produces results consistent for all samples and distributions of the innovations studied. The second study found that the Jarque-Bera test (JB) can be safely applied to a broadclass of GARCH models -M; however, it did not examine the GARCH (1, 1) models.

Science Publications

This study aims to contribute to this scarce literature by investigating how errors following a GARCH (1, 1) processes affect the statistical power and size of eight normality tests by holding out a Monte Carlo study. These tests are Shapiro-Wilk (SP), Jarque-Bera (JB), D'Agostino-Pearson, (K), Pearson (PCHI), Shapiro-Francia(SF) anderson-Darling(WCM), Lilliefors (LKS) and Cramér-von Mises (CM).

Given the literature, we have several hypotheses about the effect that heteroscedastic errors are going to have on statistical power and size of the normality test: (i) statistical power will improve as the sample size grows, (ii) the behavior of the statistical power will vary among the distributions choose in the Monte Carlo study, those that are similar to a normal distribution (such t Student) will have better statistical power and (iii) in small sample Shapiro – Wilk will be the most powerful test.

The remainder of the paper is structured as follows. In the next section, we describe the method and the data generating process. In section three, we present our findings from the Monte Carlo simulations. The paper continues with a discussion of the results in section four. At last, the conclusions and contributions of this study.

Materials and Methods

Following Alonso and Montenegro (2015), we test the behavior of eight of the most popular normality tests in the literature under the presence of heteroscedastic errors that follow a GARCH (1, 1) process. The test we study(fromnow on we will refer to them by the names in brackets) are: (i) Shapiro-Wilk (SP), (ii) Jarque-Bera (JB), (iii) D 'Agostino-Pearson, (K),(iv) Pearson (PCHI), (v) Shapiro -Francia (SF), (vi) Anderson-Darling (WCM),(vii) Lilliefors (LKS) and (viii) Cramér-von Mises (CM). We present the statistic for each test in Appendix 1.

Following Bera & Ng (1993), it is possible to classify these eight normality tests into two categories: distance tests and goodness of fit tests. The distance tests are the CM, WCM and LKS. CM is an Empirical Distribution Function (EDF) test that compares the cumulative distribution function (CDF) of a normal distribution with the estimated distribution function from the sample data and evaluates how similar are they (Razali and Wah, 2011). WCM, also an EDF test, uses the Cramérvon Mises statistic weighted with its accumulative distribution function so that the tails of the estimated distribution have more weight than the CM test. The LKS is a modification of the Kolmogorov-Smirnov (KS) test, while its statistic is determined in the same way as KS's, the critical values are not the same; therefore, LF leads to different conclusions (Razali and Wah, 2011).

On the other hand, the considered normality tests that belong to the goodness of fit tests category are JB, K, PCHI, SP and SF. JB and K are moment tests since the detection of non normality distribution come from evaluating two sample moments: skewness and kurtosis.

The main difference between those two is the transformation made to the sample moments. Besides that, both tests compare their statistic with a critical value from a Chi-Square distribution with two degrees of freedom (Singh and Masuku, 2014). The PCHI test implies a statistic that is the sum of the ratio of the squared difference between the observed frequency of data of type i and its expected frequency (Mbah and Paothong, 2014), weighted by its expected frequency. SP evaluates if a random sample comes from a normal distribution. It sums the square of the ordered sample values weighted by a constant, generated from the means, variance and covariances of the corresponding order statistics of the sample and divides the results by the sum of the square of the deviations (Mbah and Paothong, 2014). The SF is a similar test to the SP, but it is designed for large samples.

In this Monte-Carlo experiment, we consider the effect on normality tests' power of: (i) the sample size, (ii) distribution and (iii) parameter values of a GARCH (1, 1) process. Especially, the experiment will consider:

- Six sample sizes (T): 25, 50, 100, 200, 500, 1000 and 3000
- Six distributions of the error term: Standard Normal (N[0,1]), Student's t with three, five and 10 degrees of freedom, Laplace and Uniform. These last two tests come from the Generalized Error Distribution (GED) with one degree of freedom and 10 degrees of freedom, respectively
- Three types of GARCH(1,1) process: (i) $\delta = 0.1$ and $\beta=0.8$, (ii) $\beta = 0.4$ and $\delta=0.4$ and (iii) $\delta=0.8$ and $\beta=0.1$. Three types of GARCH(1,1) process: (i) $\delta = 0.1$ and $\beta=0.8$, (ii) $\delta = 0.4$ and $\beta = 0.4$ and (iii) $\delta = 0.8$ and $\beta = 0.1$. The results from these types of heteroscedastic error terms are compared with the results from a homoscedastic error term to measure how each type of heteroscedasticity affects the power and size of the normality tests

The Monte Carlo experiment implies the following steps:

1. Generate data for the vector $y_{y\times 1}$ using the following data generating process:

$$y_t = 1 + 1x_t + \varepsilon_t \# \quad (1)$$

where, x_t corresponds to a non-stochastic variable generated a priori (and only once) from a uniform distribution between zero and one. The random vector ε_t is a not auto correlated error term but is heteroscedastic and follows a GARCH (1,1) process:

$$\varepsilon_t = v_t * h_t^{\frac{1}{2}} \# \tag{2}$$

$$h_t = \omega + \delta\varepsilon_{t-1}^2 + \theta h_{t-1} \# \tag{3}$$

h_t is the conditional variance of the error and v_t is a white noise process. ω, δ and θ are constants[a]. For all cases, ω is set to 0.000001.

2. Regress y_t into x_t and a constant by ordinary least squares method, which minimize the sum of the squares of the differences between $y_{T\times 1}$ and those predicted ($\hat{y}_{T\times 1}$). In other words:

$$Min_{\hat{\beta}}\left\{\left[y_{T\times 1} - X_{T\times 1}\hat{\beta}_{2\times 1}\right]^T \left[y_{T\times 1} - X_{T\times 1}\hat{\beta}_{2\times 1}\right]\right\} \# \tag{4}$$

$$\hat{\beta}_{2\times 1} = (X_{T\times 1}{}^T X_{T\times 1})^{-1} X_{T\times 1}{}^T y_{T\times 1} \# \tag{5}$$

3. Obtain the error term:

$$\hat{\varepsilon}_{T\times 1} = \hat{y}_{T\times 1} - (X_{T\times 1})_{T\times 2}\hat{\beta}_{2\times 1} \# \tag{6}$$

4. Apply normality tests to estimated residuals ($\hat{\varepsilon}_{T\times 1}$) and record if the null hypothesis is rejected (significance level of 0.05) or not[b].
5. Repeat 10,000 times steps one throw four.
6. Calculate the observed size or power, depending on the case, as the proportions of rejections.

Results

Standard Normal Distribution

When the error is homoscedastic, normality tests show a statistical size close to nominal (α = 0.05) for all sample sizes (Table 1). In small samples, CM test has the closest value to the nominal. For samples of size 1000, the best tests is WCM. An interesting result arises when considering a heteroscedastic error term. In general, the observed statistical size of the tests becomes distorted regardless of the values of δ and β. There are a few unexpected exceptions for small samples. For example, all tests continue to show a size close to the nominal (0.05) for an error term following a GARCH(1, 1) model with δ =0.1 and β = 0.8 and sample size 25 and 50. Moreover, for δ =0.4 and β = 0.4 only for sample size 25 the statistical size of all eight tests is relatively close to 0.05. In all other cases, the observed size is far from the theoretical (see Table 1). On the other hand, for the GARCH (1, 1) model with δ =0.1 and β = 0.8, JB has the lowest empirical size among the eight tests for samples of 25 and 50 observations; however, for large samples (500,1000 and 3000) JB presents the greatest distortion. For samples between 100 and 3000 observations, the PCHI presents the lowest statistical size compared to the other eight tests; despite that, as the number of observations growth statistical size also grows and becomes 0.103, twice the nominal. For the GARCH (1, 1) model with δ =0.4 and β = 0.4 we obtain similar results. One important difference for samples of size1000 from the previous case is that all tests exhibit on average a probability of96.6% (disregarding the PCHI that has the smallest size) of making the mistake of rejecting the null hypothesis when it is true. Finally, all tests, except CM, present the greatest distortions under errors from a GARCH (1, 1) model with δ =0.8 and β = 0.1, since for samples of 500 or more observations the probability of making type I error is of 100%. Power with Error Term from Student's t- Distribution

Tables 2, 3 and 4 present results for Student's t distribution with three, five and 10 degrees of freedom, respectively. In general terms, the power of normality tests is about one when the sample size is 1000 or 3000; except the CM test that has the lowest power in those two samples sizes. However, the power decreases as we increase the number of degrees of freedom because the t-distribution approaches a normal distribution. This phenomenon intensifies in samples of size 25 and 50, but it is almost imperceptible in large samples. For example, for the homoscedastic residuals, the SF test has the greatest power in the distribution with three degrees of freedom, for five degrees its power reduces to 0.239 and for the distribution with 10 degrees of freedom is 0.124. For heteroscedastic residuals, the power of the same test is 0.387, 0.199 and finally 0.096, respectively. On the other hand, the SF test has the highest empirical power for all samples considered when the error has a constant variance and follows a distribution with three degrees of freedom. The same applies to Student's t-distribution with five degrees of freedom, excluding the sample of 25 observations. Instead, for a Student's t- distribution with 10 degrees of freedom, JB shows the best power for samples of 100 to 3000 observations. The SF is the most powerful test under the three types of heteroscedasticity and for the three Student's t-distributions considered. It is interesting to note in Table 3 and 4 that GARCH (1, 1) model with δ =0.4 and β = 0.4 and δ =0.8 have a positive effect on the test's power when compare with the power obtained from applying the tests to the homoscedastic case. However, the above does not hold for the CM test when it is applied to samples of 1000 and 3000 observations since the probability of rejecting a false hypothesis is significantly reduced. For example, it becomes 0% in the case of a distribution with 10 degrees of freedom and a sample of 3000 observations. Moreover, β = 0.1 have a positive effect on the test's power when compare with the power obtained from applying the tests to the homoscedastic case. Power with Error Term from a Laplace Distribution Results for the Laplace distribution are similar to those of the Student's t distribution. Tests show a statistical power close or equal to one for the homoscedastic and heteroscedastic residuals in samples of 500, 1000 and 3000 observations (see Table 5). For small sample sizes, all tests have relatively low power. That improves when errors come from GARCH(1,1) models with δ =0.4 and β = 0.4 and δ =0.8 and β = 0.1. Moreover, when δ =0.1 and β = 0.8 empirical power for all tests is worse. SF has the biggest power for samples of size 25, 50 and 100 in both heteroscedastic and homoscedastic error. For those same cases, PCHI has the lowest power.

Power with Error Term from a Uniform Distribution

Results for this distribution are similar to those found with the Student's t and Laplace distribution (see Table 6). All tests have power equal or close to one in large samples (500, 1000 and 3000 observations). The CM is the only test that shows a statistical power of 0% for a sample of 3000 when the error has a constant variance over time or comes from GARCH(1,1) models with δ = 0.1 and β = 0.8. For those two cases, the K test shows the best power in samples of 50 and 100 observations. For δ = β = 0.4 the WCM presents the greatest power in samples of size 25, 500, 1000 and 3000; however, only for the last two sample sizes, the statistical power is above 0.9. Finally, for from δ = 0.8 and β = 0.1, SF is the test with the best power in samples of 50 and 100, as in the Laplace distribution and the power is greater than 0.9 in large samples. Furthermore, when δ=0.1 and β=0. 8 normality test's statistical power increases slightly in samples of 25 observations, but it decreases in samples of 50 and 100 observations and for large samples, there is no distortion.

When δ=β=0.4, statistical powered creases for all sample sizes (except 3000) in comparison with homoscedastic errors. Finally, when δ =0.8 and β = 0.1 the statistical power improves in relation to the case when δ = β = 0.4. However, the power is still less than the one obtained when δ =0.1 and β = 0.8 and the homoscedastic case.

Table 1: Statistical size of normality tests under errors following a standard normal distribution

		Normality test							
Garch (1,1) Model	Sample (S)	SP	JB	K	PCHI	SF	WCM	LKS	CM
Homoscedastic	25	0,045	0,026	0,054	0,057	0,048	0,047	0,049	0,049
	50	0,046	0,034	0,055	0,048	0,047	0,047	0,051	0,050
	100	0,052	0,042	0,055	0,051	0,054	0,051	0,049	0,050
	500	0,048	0,043	0,047	0,053	0,049	0,048	0,047	0,049
	1000	0,051	0,049	0,049	0,051	0,054	0,050	0,049	0,052
	3000	0,047	0,050	0,051	0,052	0,052	0,044	0,039	0,044
δ=0.1 β=0.8	25	0,043	0,021	0,049	0,059	0,042	0,045	0,043	0,045
	50	0,051	0,036	0,058	0,053	0,053	0,049	0,048	0,050
	100	0,070	0,074	0,087	0,054	0,081	0,065	0,058	0,063
	500	0,184	0,243	0,212	0,062	0,222	0,120	0,081	0,105
	1000	0,308	0,378	0,342	0,067	0,352	0,188	0,108	0,160
	3000	0,624	0,728	0,698	0,103	0,687	0,420	0,218	0,350
δ=0.4 β=0.4	25	0,075	0,054	0,094	0,073	0,087	0,073	0,061	0,071
	50	0,162	0,171	0,188	0,082	0,196	0,142	0,109	0,131
	100	0,332	0,376	0,359	0,134	0,384	0,284	0,209	0,258
	500	0,882	0,915	0,895	0,404	0,904	0,806	0,657	0,751
	1000	0,988	0,994	0,992	0,665	0,992	0,972	0,894	0,928
	3000	1	1	1	0,988	1	1	1	0,814
δ=0.8 β=0.1	25	0,171	0,152	0,200	0,118	0,208	0,170	0,135	0,163
	50	0,403	0,420	0,426	0,215	0,451	0,368	0,291	0,346
	100	0,691	0,730	0,709	0,398	0,733	0,643	0,530	0,603
	500	0,999	1	1	0,936	1	0,997	0,988	0,732
	1000	1	1	1	0,998	1	1	1	0,358
	3000	1	1	1	1	1	0,968	1	0

Note: The closest statistical size to the theoretical size (0.05) for each sample and type of error is in bold.

Table 2: Statistical power of normality tests under errors following a t-distribution with 3 degrees of freedom

		Normality Test							
Garch (1,1) Model	Sample (S)	SP	JB	K	PCHI	SF	WCM	LKS	CM
Homoscedastic	25	0.373	0.363	0.419	0.202	0.424	0.348	0.266	0.323
	50	0.620	0.644	0.645	0.306	0.678	0.581	0.455	0.545
	100	0.873	0.891	0.872	0.520	0.903	0.847	0.724	0.816
	500	1	1	1	1	1	1	1	0.847
	1000	1	1	1	1	1	1	1	0.336
	3000	1	1	1	1	1	0.997	1	0
δ=0.1 β=0.8	25	0.335	0.326	0.383	0.167	0.387	0.312	0.231	0.289
	50	0.601	0.627	0.624	0.292	0.659	0.565	0.439	0.530
	100	0.866	0.886	0.865	0.525	0.897	0.840	0.725	0.814
	500	1	1	1	1	1	1	1	0.755
	1000	1	1	1	1	1	1	1	0.202
	3000	1	1	1	1	1	0.987	1	0
δ=0.4 β=0.4	25	0.358	0.341	0.402	0.196	0.411	0.341	0.261	0.320
	50	0.632	0.649	0.648	0.358	0.688	0.607	0.494	0.581
	100	0.892	0.906	0.888	0.610	0.916	0.875	0.785	0.842
	500	1	1	1	1	1	1	1	0.557
	1000	1	1	1	1	1	1	1	0.058
	3000	1	1	1	1	1	0.961	1	0
δ=0.8 β=0.1	25	0.402	0.379	0.436	0.240	0.453	0.391	0.311	0.373
	50	0.686	0.700	0.698	0.431	0.738	0.671	0.563	0.645
	100	0.923	0.933	0.918	0.706	0.942	0.914	0.846	0.868
	500	1	1	1	1	1	1	1	0.328
	1000	1	1	1	1	1	1	1	0.328
	3000	1	1	1	1	1	0.999	1	0.008

Note: The highest statistical power for each sample and type of error is in bold.

Table 3: Statistical power of normality tests under errors following a t-distribution with 5 degrees of freedom

		Normality Test							
Garch (1,1) Model	Sample (S)	SP	JB	K	PCHI	SF	WCM	LKS	CM
Homoscedastic	25	0.199	0.188	0.242	0.104	0.239	0.176	0.122	0.156
	50	0.341	0.379	0.387	0.121	0.400	0.289	0.195	0.256
	100	0.548	0.617	0.587	0.179	0.620	0.466	0.321	0.417
	500	0.993	0.995	0.993	0.578	0.995	0.979	0.898	0.966
	1000	1	1	1	0.896	1	1	0.997	0.997
	3000	1	1	1	1	1	1	1	0.774
δ=0.1 β=0.8	25	0.164	0.158	0.205	0.095	0.199	0.145	0.108	0.133
	50	0.310	0.344	0.357	0.118	0.371	0.270	0.190	0.240
	100	0.558	0.616	0.586	0.208	0.624	0.494	0.356	0.449
	500	0.996	0.996	0.995	0.729	0.997	0.990	0.944	0.972
	1000	1	1	1	0.966	1	1	0.999	0.958
	3000	1	1	1	1	1	1	1	0.219
δ=0.4 β=0.4	25	0.202	0.188	0.238	0.120	0.241	0.195	0.147	0.180
	50	0.418	0.443	0.453	0.197	0.479	0.385	0.293	0.359
	100	0.701	0.739	0.711	0.368	0.755	0.659	0.535	0.627
	500	1	1	1	0.938	1	0.999	0.992	0.881
	1000	1	1	1	0.9983	1	1	1	0.573
	3000	1	1	1	1	1	0.994	1	0.0002
δ=0.8 β=0.1	25	0.290	0.267	0.325	0.173	0.338	0.279	0.223	0.268
	50	0.558	0.576	0.579	0.315	0.613	0.537	0.439	0.509
	100	0.840	0.858	0.837	0.563	0.871	0.816	0.723	0.774
	500	1	1	1	0.993	1	1	1	0.547
	1000	1	1	1	1	1	0.999	1	0.086
	3000	1	1	1	1	1	0.930	1	0

Note: The highest statistical power for each sample and type of error is in bold

Table 4: Statistical power of normality tests under errors following a t-distribution with 10 degrees of freedom

		Normality Test							
Garch (1,1) Model	Sample (S)	SP	JB	K	PCHI	SF	WCM	LKS	CM
Homoscedastic	25	0.105	0.089	0.131	0.076	0.124	0.091	0.070	0.084
	50	0.151	0.170	0.184	0.063	0.186	0.118	0.084	0.104
	100	0.225	0.283	0.268	0.070	0.278	0.157	0.107	0.136
	500	0.652	0.743	0.692	0.123	0.713	0.478	0.275	0.408
	1000	0.900	0.944	0.923	0.198	0.926	0.781	0.498	0.703
	3000	1	1	1	0.605	1	0.997	0.950	0.993
δ=0.1 β=0.8	25	0.082	0.063	0.100	0.072	0.096	0.074	0.061	0.070
	50	0.138	0.156	0.172	0.066	0.171	0.109	0.081	0.100
	100	0.254	0.309	0.289	0.087	0.311	0.200	0.132	0.176
	500	0.809	0.863	0.827	0.225	0.849	0.700	0.486	0.640
	1000	0.972	0.984	0.977	0.428	0.980	0.939	0.785	0.909
	3000	1	1	1	0.931	1	1	0.998	0.998
δ=0.4 β=0.4	25	0.127	0.107	0.148	0.083	0.151	0.117	0.092	0.110
	50	0.273	0.291	0.303	0.122	0.323	0.245	0.183	0.226
	100	0.512	0.558	0.532	0.220	0.566	0.458	0.344	0.425
	500	1	1	1	0.710	1	1	0.915	0.926
	1000	1	1	1	1	1	1	1	0.888
	3000	1	1	1	1	1	1	1	0
δ=0.8 β=0.1	25	0.223	0.203	0.253	0.135	0.263	0.217	0.171	0.211
	50	0.478	0.495	0.498	0.258	0.530	0.449	0.357	0.425
	100	0.760	0.793	0.769	0.469	0.801	0.727	0.621	0.688
	500	1	1	1	0.975	1	1	0.997	0.662
	1000	1	1	1	1	1	1	1	0.220
	3000	1	1	1	1	1	0.953	1	0

Note: The highest statistical power for each sample and type of error is in bold

Table 5: Statistical power of normality tests under errors following a Laplace distribution

		Normality Test							
Garch (1,1) Model	Sample (S)	SP	JB	K	PCHI	SF	WCM	LKS	CM
Homoscedastic	25	0.268	0.250	0.315	0.151	0.334	0.268	0.207	0.264
	50	0.482	0.487	0.486	0.237	0.561	0.497	0.386	0.49
	100	0.775	0.764	0.718	0.432	0.828	0.803	0.677	0.795
	500	1	1	1	0.993	1	1	1	1
	1000	1	1	1	1	1	1	1	0.948
	3000	1	1	1	1	1	1	1	0
δ=0.1	25	0.227	0.202	0.264	0.135	0.286	0.232	0.178	0.224
β=0.8	50	0.458	0.456	0.452	0.231	0.534	0.477	0.374	0.469
	100	0.785	0.770	0.726	0.461	0.831	0.81	0.691	0.808
	500	1	1	1	0.998	1	1	1	0.986
	1000	1	1	1	1	1	1	1	0.66
	3000	1	1	1	1	1	1	1	0
δ=0.4	25	0.286	0.250	0.313	0.177	0.346	0.293	0.233	0.284
β=0.4	50	0.567	0.561	0.558	0.326	0.635	0.596	0.48	0.578
	100	0.871	0.857	0.8256	0.621	0.902	0.887	0.808	0.881
	500	1	1	1	1	1	1	1	0.744
	1000	1	1	1	1	1	1	1	0.096
	3000	1	1	1	1	1	0.995	1	0
δ=0.8	25	0.373	0.344	0.402	0.232	0.437	0.378	0.311	0.368
β=0.1	50	0.687	0.678	0.677	0.45	0.739	0.695	0.595	0.685
	100	0.935	0.929	0.91	0.767	0.951	0.945	0.894	0.919
	500	1	1	1	1	1	1	1	0.317
	1000	1	1	1	1	1	1	1	0.003
	3000	1	1	1	1	1	0.915	1	0

Note: The highest statistical power for each sample and type of error is in bold

Table 6: Statistical power of normality tests under errors following a Uniform distribution

		Normality Test							
Garch (1,1) Model	Sample (S)	SP	JB	K	PCHI	SF	WCM	LKS	CM
Homoscedastic	25	0.149	0.001	0.150	0.103	0.057	0.138	0.085	0.121
	50	0.456	0.000	0.587	0.161	0.232	0.384	0.192	0.311
	100	0.911	0.271	0.971	0.365	0.763	0.817	0.466	0.696
	500	1	1	1	1	1	1	1	1
	1000	1	1	1	1	1	1	1	1
	3000	1	1	1	1	1	1	1	0
δ=0.1	25	0.154	0.001	0.151	0.105	0.066	0.148	0.097	0.133
β=0.8	50	0.404	0.000	0.547	0.166	0.209	0.360	0.186	0.303
	100	0.822	0.185	0.927	0.344	0.636	0.740	0.429	0.639
	500	1	1	1	0.998	1	1	0.999	1
	1000	1	1	1	1	1	1	1	1
	3000	1	1	1	1	1	1	0.024	δ=0.4
β=0.4	25	0.098	0.004	0.089	0.097	0.049	0.103	0.084	0.101
	50	0.155	0.009	0.216	0.109	0.082	0.169	0.121	0.160
	100	0.280	0.043	0.360	0.180	0.192	0.314	0.215	0.294
	500	0.806	0.483	0.582	0.617	0.796	0.865	0.701	0.820
	1000	0.973	0.585	0.635	0.897	0.973	0.989	0.918	0.974
	3000	1	0.671	0.686	1	1	1	1	1
δ=0.8	25	0.101	0.057	0.104	0.097	0.102	0.111	0.094	0.110
β=0.1	50	0.232	0.217	0.248	0.139	0.252	0.220	0.172	0.204
	100	0.442	0.457	0.455	0.241	0.477	0.393	0.303	0.354
	500	0.962	0.962	0.956	0.733	0.969	0.920	0.810	0.753
	1000	0.999	0.999	0.999	0.954	0.999	0.997	0.974	0.738
	3000	1	1	1	1	1	0.988	1	0.200

Note: The highest statistical power for each sample and type of error is in bold

Discussion

Concerning the hypothesis presented in the introduction, results show that almost all of what we have stated occur with the data. First, the statistical power improves as the sample size grows. For samples between 500 and 3000 observations, any test can be used because they exhibit good power; while in small samples is preferable to use Shapiro-Francia because it has the highest power when compared to the other tests.

Second, the behavior of the statistical power varied among the distributions chosein the Monte Carlo study. The power of the tests increases or decrease depending on the type of heteroscedasticity and distribution considered. This effect is greater in medium- size samples (50, 100 and 500 observations). Contrary of what we thought, the statistical power of the normality test under a Student's t distribution was not the least affected by the types of heteroscedasticity considered; it was under the Laplace distribution that normality tests were not substantially affected.

Besides that, the power of all test was low for small samples. This result has also been found in other studies where the error term does not meet all the assumptions. For example, Alonso and Montenegro (2015) evaluated normality tests in the presence of errors that follow an AR (1) process. Results showed that the effect of autocorrelation on the power of the tests is asymmetrical, the statistical is distorted inthe presence of strong autocorrelation and all tests have a similar power, which tends to be low for small samples.

Third, not always the Shapiro –Wilk test was the most powerful test in small samples. Instead, Shapiro - Francia has a better power in a small sample. Similar results have been found in simulations about the performance of the normality test under other conditions. Razali and Wah (2011) studied the power of four normality tests (Shapiro and Wilk (1965), Kolmogorov (1933), Lilliefors (1967) and Anderson and arling (1952)) for symmetric and asymmetric distributions and 15 different sample sizes. They concluded that Shapiro-Wilk is the most powerful test, followed by Anderson-Darling, Lilliefors and finally Kolmogorov-Smirnov. The two last tests required a sample size close to 2000 observations to obtain a power likethe Shapiro-Wilk test.

Mbah and Paothong (2014) compared the performance of the Shapiro-Francia test with other eight normality test by studying the distribution of their p-values. They found that Shapiro-Francia is the best test for detecting deviations from normality from the eight tests analyzed. The Monte Carlo simulation set up consisted of 12 sample sizes and eight distributions for the error term (not correlated and homoscedastic), some of them were the standard normal distribution, uniform, Laplace and Student-t distribution with different degrees of freedom.

Future research will include other normality tests, such as QH*, which was proposed by Chen and Shapiro (1995). This test had had a better performance than other normality tests based on regression under a diverse combination of symmetric, asymmetric, contaminated and balanced distributions and samples size of 20, 50 and 100 observations (Seier, 2002). Also, it will involve developing and proposing those statistics for each normality test, under each variation of GARCH (1,1) model and distribution analyzed, to capture the characteristics the sample must have to have a theoretical size of 0.05 while maximizing the power of the test.

Conclusion

This paper has presented a Monte Carlo study that describes how the power and size of eight normality test behave under three variations of GARCH (1, 1) models for different sample sizes and distributions. Three important results are obtained from these simulations. First, the probability of making type I error, especially in samples of size 500, 1000 and 3000, increased under the presence of heteroscedastic error terms. Our results imply that Pearson's test (1900) is a suitable choice for samples of size 25, 50 and 100. For larger samples, our results suggest being cautious and complement the validation of normality assumption with other tools since using any of the eight tests studied could lead to wrong conclusions.

Second, in the homoscedastic case, we should apply the Cramér-von Mises. In the presence of GARCH (1, 1) the Pearson (1900) test should be use. Third, the normality tests have low statistical power and size (less than or equal to the nominal employee ($\alpha = 0.05$)), in small and homoscedastic samples.

Fourth, the recommendation of our experimental work to the scientific community is to use other tools besides the normality tests for making an inference from small samples and for constructing prediction intervals of any sample size. Nonparametric approaches should be considered.

Future research should focus on designing appropriate normality tests when the error term follows a GARCH (1,1). An idea that is worth evaluating is the statistics proposed by Jarque and Bera (1980). Jarque and Bera (1980), unlike Jarque and Bera (1987), proposed a test for normality accounting for heteroscedasticity and serially correlation. This statistic has not been adapted for the problems addressed in this paper. For simplicity, practitioners use Jarque and Bera (1987) test that is a simplified version of the Jarque and Bera (1980) that does not account for heteroscedasticity and autocorrelation. The Jarque and Bera (1980) test needs that the researcher specifies the form of the covariance matrix of the error term and is not implemented in commercial software. This may be a good starting point to design a better test of normality under GARCH (1,1) behavior.

Funding Information

No financial support was received, neither technical assistance.

Author's Contributions

Authors contributed to the same extent to all the process of preparing and developing the manuscript since we operate as a group.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and there are no ethical issues involved.

References

Alonso, J. and S. Montenegro, 2015. Estudio de Monte Carlo para comparar 8 pruebas de normalidad sobre residuos de mínimos cuadrados ordinarios en presencia de procesos autoregresivos de primer orden. EstudiosGerenciales.

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

Bellini, F. and L. Bottolo, 2008. Misspecification and domain issues in fitting Garch(1, 1) models: A monte carlo investigation. Communications Statistics Simulation Computation, 38: 31-45.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econometrics, 31: 307-327.

Bollerslev, T., R. Chou and K. Kroner, 1992. ARCH modeling in finance. A review of the theory an empirical evidence. J. Econometrics, 52: 5-59.

Chen, L. and S.S. Shapiro, 1995. An alernative test for normality based on normalized spacings. J. Statistical Computation Simulation, 53: 269-287.

Enders, W., 2003. Modeling Volatility. In: Applied Econometric Time Series, Wiley, pp: 108-155.

Engle, R. and G. Gonzalez-Rivera, 1991. Semiparametric ARCH models. J. Bus. Economic, 9: 345-359.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom Inflation. Econometrica, 50: 987-1007.

Fiorentini, G., E. Sentana and G. Calzolari, 2004. On the validity of the Jarque-Bera normality test in conditionally heteroskedastic dynamic regression models. Economics Lett., 83: 307-312.

Green, W.H., 2012. The Generalized Regression Model and Heterocedasticity. In: Econometric Analysis, Pearson, pp: 257-289.

Jarque, C.M. and A.K. Bera, 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters, 6: 255-259.

Jarque, C.M. and A.K. Bera, 1987. A test for normality of observations and regression residuals. Int. Statistical Review/Revue Internationale de Statistique, pp: 163-172.

Klar, B., F. Lindner and S.G. Meintanis, 2012. Specification tests for the error distribution in GARCH models. Computational Statistics Data Analysis, 56: 3587-3598.

Kolmogorov, A.N., 1933. Sulla determinazione empirica di une legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari, 4: 83-91.

Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Statistical Association, 62: 399-402.

Mbah, A.K. and A. Paothong, 2014. Shapiro Francia test compared to other normality test using expected p - value. J. Statistical Computation Simulation.

Razali, N.M. and Y.B. Wah, 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J. Statistical Modeling Analytics, 2: 21-33.

Seier, E., 2002. Comparison of tests for univariate normality tests for normality. Interstat.

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

Singh, A.S. and M.B. Masuku, 2014. Assumption and testing of normality for statistical analysis. Am. J. Math. Mathematical Sci., 3: 169-175.

Vavra, M., 2011. Testing normality in time series. Phd Thesis of Philosophy, Birkbeck College, University of London, United Kingdom.

Appendix 1: Statistical for the eight considered tests. Using a sample of X_i random selected elements, lets define the following statistics and quantities:

- s: sample's standard deviation
- S: sample's skewness
- K: sample's kurtosis

The Jarque-Bera (JB) statistical is defined as:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right)\# \tag{7}$$

And D 'Agostino-Pearson (DA) is:

$$DA = Z^2(S) + Z^2(K)\# \tag{8}$$

where, Z^2 (·) is the squared standard normal distribution. For the following normality tests, lets define the next vectors and matrices:

- **X**: vector of dimensions $(1\times n$)containing the order statistic of the sample ($XX(i$))
- σ^2**V**: The covariance matrix of the vector of all ($X_{(i)}$)
- *c*: vector of expected values of the n order statistics from a normal standard distribution
- **a**: vector of dimensions $(1\times n$) such that:

$$a' = \frac{c'V^{-1}}{(c'V^{-2}c)^{1/2}}\# \tag{8}$$

- b: vector of dimensions $(1\times n$) such that:

$$b' = \frac{c'}{(c'c)^{1/2}}\# \tag{9}$$

Then the Shapiro-Wilk (SP) statistic is defined as:

$$SP = \frac{(a'X)^2}{(n-1)s^2}\# \tag{10}$$

And the Shapiro-Francia (SF):

$$SF = \frac{(b'X)^2}{n\hat{\sigma}^2} = \frac{\left(\sum_{i=1}^{n} b_i\, x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\# \tag{11}$$

Let's define the following variables:

- *k*: number of classes in the sample
- O_i: observed frequency of the ith bin
- E_i: expectation calculated as E_i=n($F(X_u)-F(X_l)$) where X_u and X_l are the lower and upper bounds of the ith bin, respectively and n is the sample size

Then, the Pearson (PCHI) is:

$$x^2\sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i}\# \tag{12}$$

Finally, let's the define the following function and statistic:

- *Z*: normal density function with mean and variance unknown for i = 1, 2, ···, *n*
- *KS*: Kolmogorov-Smirnov statistic which is defined as KS ={$KS+$,KS^-}, where KS^+=[(i /n)−z_i] and $KS - [z_i - \frac{(i-1)}{n}]$, with $1\leq i \leq n$.

Then Anderson-Darling (WCM) is defined as:

$$WCM = -n - \frac{1}{n}\sum_{i=1}^{n}[(2i-1)(In\ z_i + In(1-z_{n+1-i}))] \qquad (13)$$

And the Cramér-von Mises (CM) is:

$$CM = \frac{1}{12n} + i = 1n\left(z_i\ \frac{2i-1}{2n}\right)^2 \# \qquad (14)$$

At last, Lilliefors (LKS) is defined as:

$$LKS = \begin{cases} KS, n \leq 100 \\ KS\left(\frac{n}{100}\right)^{0.49}, n > 100 \end{cases}$$

[a] δ and θ are two parameters that increase the conditional volatility, but they do it in different ways. The larger is δ, the greater the response of h_t to new information; if δ is large, then a shock of v_t affects ε_t and h_{t+1} then after the shock effect would be observed pronounced in the variance of the next period. In contrast, the greater is θ the conditional variance show a persistent autoregressive process and correspond to more permanent peaks. To guarantee variance convergence, non-negative variance and stationary $\delta+\theta$ must be <1 and for a non-negative variance $\omega\omega$) must be > 0, $\delta \leq 0$ and $\theta \leq 0$.

[b] For all eight normality tests considered, the null hypothesis is that the sample comes from a normal distribution and the alternative is that it originates from a non-normal distribution. The null hypothesis is rejected if the p-value associated with the test statistics is lesser than 0.05, which the significance level.