Macroeconomic Dynamics, **14** (Supplement 1), 2010, 137–144. Printed in the United States of America.
doi:10.1017/S1365100510000015

A TEST OF THE GARCH(1, 1) SPECIFICATION FOR DAILY STOCK RETURNS

RICHARD A. ASHLEY AND DOUGLAS M. PATTERSON
Virginia Tech (VPI)

Daily financial returns (and daily stock returns, in particular) are commonly modeled as GARCH(1, 1) processes. Here we test this specification using new model evaluation technology developed by Ashley and Patterson that examines the ability of the estimated model to reproduce features of particular interest: various aspects of nonlinear serial dependence, in the present instance. Using daily returns to the CRSP equally weighted stock index, we find that the GARCH(1, 1) specification cannot be rejected; thus, this model appears to be reasonably adequate in terms of reproducing the kinds of nonlinear serial dependence addressed by the battery of nonlinearity tests used here.

Keywords: GARCH, Stock Returns, Volatility, Model Evaluation

1. INTRODUCTION

The nonlinear serially dependent ARCH/GARCH and EGRACH group of models are widely accepted among econometricians and time series statisticians as the premier model of stock market returns, especially so for the GARCH(1, 1) model [see Bollerslev (1986) for the derivation of the model]. This wide acceptance rests on two bodies of empirical evidence. First, a number of statistical tests (bispectrum, BDS, Engel's LM, etc.) easily reject the null hypothesis of a linear process; this evidence against a linear process has been accumulating since the mid-1980s. [For example, see Hinich and Patterson (1985) and Patterson and Ashley (2000).] Second, the parameter estimates of a GARCH(1, 1) process are statistically significant when a model is estimated on various examples of realized stock market returns—market indices and individual stock issues. This statistical significance of the parameter estimates is apparently sufficient evidence for the vast majority of empirical investigators to accept these models as true. In the paper "ARCH Modeling in Finance," Bollerslev et al. (1992) cite over 300 papers, most of which touch at some point on the ARCH family of models to empirically study either the volatility or the risk premium in asset returns.

On the other hand, there is not a consensus among finance academics that the ARCH/GARCH specification for stock market returns dominates all other

Address correspondence to: Douglas M. Patterson, Department of Finance (0221), Virginia Tech, Blacksburg, VA 24061, USA; e-mail: amex@vt.edu.

 1365-1005/10

models. One reason for this is that there is no compelling theory to explain why returns should be generated by this particular family of nonlinear processes. In addition, although it is generally agreed that market volatility varies over time, there are alternative (and linear) models of returns that can also explain time-varying volatility—e.g., stochastic volatility models, seasonal ARIMA models, and random jump processes.

2. MODEL EVALUATION

In this paper we carefully evaluate the ability of the ARCH family of models to explain the nonlinear dependence of stock market returns. A common approach for comparing alternative time series models is to ask which model fits the data best, based on R_c^2, FPE, AIC, BIC, etc. Sample fit is important, but because the sample data are customarily (and necessarily) mined to identify the particular form of whatever kind of model is being considered, the fact that the resulting model fits the data well usually reflects the flexibility of the framework being used (GARCH, threshold autoregressive, Markov switching, neural net, or whatever) more than it does which kind of model is closer to the specification that actually generated the data.

Another approach relies on relative out-of-sample forecasting effectiveness as a criterion for model choice. [See, for example, Diebold and Mariano (1995), Ashley (1998), and McCracken (2007).] Out-of-sample forecasting can give substantially credible support to a particular model or to one model specification over another. But the results from this approach can be idiosyncratic to the particular model validation period chosen unless the holdout sample is lengthy, in which case an insufficient number of observations may remain for model specification and estimation. (In particular, one might expect that an adequate postsample forecast period for evaluating a state-switching model would need to be sufficiently long to include a number of state switches.) Quite often, moreover, one finds that neither of two candidate nonlinear time series models provides out-of-sample forecasts that are very useful; in such cases it seems unreasonable to prefer one model to the other on this basis. Such poor out-of-sample forecasting can arise because both model specifications are totally inadequate, but it can also reflect the fact that forecasts from nonlinear models are very sensitive to even modest model misspecification. In other words, it might be the case that one model is substantially closer to the true data-generating mechanism in the ways we most care about, yet neither model is close enough to forecast out of sample creditably well.

3. A NEW APPROACH

In Ashley and Patterson (2006) we introduced a new approach—complementary to the "sample fit" and "out-of-sample forecasting" approaches outlined above—for either evaluating an individual nonlinear times series model or comparing two such models. Our approach is based on a battery of distinct nonlinearity tests.

The reason that there are so many tests for nonlinear serial dependence (and the reason that no comprehensive model identification algorithm for nonlinear models has found widespread acceptance) is that there are many distinctly different ways in which the current value of a time series can depend nonlinearly on its own past. Consequently, many tests for nonlinearity can be constructed, each focusing on a different aspect or effect of nonlinear serial dependence—e.g., one test might focus on the way nonlinear serial dependence impacts the higher-order moments of the time series, whereas another test might look at how close different sequential m-tuples of the process are to each other. Thus, some nonlinearity tests will naturally be substantially more powerful than others against specific alternatives.

Our approach leverages this diversity by taking the pattern of p-values with which a set of nonlinearity tests rejects the null hypothesis (of a linear generating mechanism for a particular time series) as a new stylized fact characterizing the nonlinear serial dependence in this time series. One can then ask of any estimated model for this time series, "How well does it reproduce this stylized fact?"

Thus, our approach is similar in spirit to the more descriptive examination by Harding and Pagan (2002) of how well a statistical model is able to track specific features of the shape of the business cycle. Indeed, if one includes explicitly shape-related tests—e.g., the tests for steepness and depth proposed by Ramsey and Rothman (1996), Verbrugge (1997), and others—in the set of nonlinearity tests considered, then our approach subsumes and extends theirs.

One could simulate data from the estimated model and compute the power of each nonlinearity test to reject the null hypothesis of a linear generating mechanism against this particular alternative generating mechanism. If the estimated model were effective at modeling the nonlinear serial dependence in the actual data, then one would expect that the tests that were most powerful in detecting this alternative would the ones that rejected the null hypothesis with the lowest p-values using these simulated data sets. In contrast, if the tests that provided the strongest evidence for nonlinearity were ones with relatively small power to detect the kind of nonlinearities generated by this model, it seems less plausible that the actual generating mechanism for these data would be of this kind.

Our approach takes this reasoning one step further, allowing us to construct a straightforward statistical test of the proposition that a specific nonlinear model is capturing the nonlinear serial dependence in the data, as distinct from merely fitting the sample data well in a least-squares sense.

Suppose that r nonlinearity tests have been applied to the sample data, yielding r p-values $(p_1^{\text{obs}}, \ldots, p_r^{\text{obs}})$ for rejection of the null hypothesis of a linear generating mechanism for the time series. Consider, then, a "portmanteau" test statistic quantifying the discrepancy between this set of results and the set of p-values one might expect had the sample data been generated by this specific model:

$$AP(p_1, \ldots, p_r) = \sum_{i=1}^{r} [p_i - E\{p_i\}]^2.$$

Note that the expectation in this expression is taken over the joint distribution of the r-vector $(p_1, \ldots, p_r)$; this vector is a random variable because the p-value for each of the r nonlinearity tests is a monotonic transformation of the test statistic for that particular nonlinearity test.

Both this expectation and the sampling distribution of the AP test statistic are readily obtained by Monte Carlo simulation under the null hypothesis that the sample data are generated by a particular model. Indeed, these simulations are exactly those that one would do in order to calculate the power of the individual nonlinearity tests for such a model, so that the value of $E\{p_1\}, \ldots, E\{p_r\}$ obtained in this way is closely related to the usual estimate of the power of these nonlinearity tests against this alternative.[1]

The p-value at which one can reject the null hypothesis that the sample data are generated by this particular model is thus the fraction of these 1,000 simulations that yield AP values in excess of $AP(p_1^{\text{obs}}, \ldots, p_r^{\text{obs}})$. One could interpret this p-value as quantifying how unlikely it would be—under the null hypothesis that this model generated the data—to observe a sample pattern of nonlinearity test p-value results this distinct from what the powers of the individual tests against this particular alternative would suggest.

4. THE DATA

The data used for the test are the daily returns (including dividends) to the CRSP equally weighted stock index; this index includes all NYSE and AMEX and the major NASDAQ stocks. The sample period used here is January 6, 2006, through December 31, 2007, for a total of 500 daily observations.

5. NONLINEARITY TESTS CONSIDERED

The six nonlinearity tests employed here are listed in Table 1. These tests are completely documented in Patterson and Ashley (2000) and in Ashley and Patterson (2006). However, for completeness, we note here that the McLeod/Li test is implemented using 24 squared serial correlation terms; the Engle Lagrange multiplier (LM) test is implemented using five lags of the squared series; the BDS test is implemented with an embedding dimension of two and the parameter

TABLE 1. Nonlinearity tests employed

Test	Focus
McLeod/Li	ARCH/GARCH
Engle Lagrange multiplier	ARCH/GARCH
BDS	General serial dependence
Hinich bicovariance	Third-order moments (time domain)
Tsay	Quadratic terms (time domain)
Hinich bispectrum	Third-order moments (frequency domain)

ε is set to one; the Hinich bicovariance test is implemented using lags up to 500^4; the Tsay test is implemented using lags up to five; and the Hinich bispectrum test is implemented using the interdecile range dispersion measure and smoothing constant $M = 500^6$.[2] Critical points for each test are obtained via bootstrap simulation (with 10,000 replications) applied to data that were prewhitened using an AR(p) model, where p is chosen by minimizing the BIC criterion.

Many other tests for nonlinear serial dependence have been described in the literature, including Ramsey (1969), Ashley and Patterson (1986), Saikkonen and Luukkonen (1988), White (1989), Mizrach (1991), Nychka et al. (1992), Kaplan (1993), Dalle Molle and Hinich (1995), and Hansen (1999). Because asymmetry is a common consequence of nonlinear serial dependence, one might also consider tests for steepness or depth, as in Ramsey and Rothman (1996) and Verbrugge (1997). No representation is made here that the group of tests listed in the table is in any sense optimal or that these tests in any well-defined sense span the space of all possible nonlinearity tests.

6. STATISTICAL RESULTS

We considered the three members of the ARCH family that have been most commonly suggested as models of stock market returns: ARCH, GARCH, and EGARCH. Model estimation was carried out using PROC AUTOREG in the SAS system. The Lagrange Multiplier (LM) test for ARCH disturbances indicated significant dependence out to at least lag 12; therefore an ARCH model is not appropriate for these data, but rather a GARCH or EGARCH model is suggested. We attempted to fit an EGARCH(2, 1) model but the SAS algorithm did not converge. The GARCH(1, 1) model was the only specification for which the estimation procedure converged to statistically significant parameter estimates consistent with a stable model. We display significant parameter estimates for the GARCH(1, 1) model in Table 2, using the usual notation for the parameters.

We simulated this estimated GARCH(1, 1) model 1,000 times using the parameter values given in Table 2. After the first 300 generated data values were dropped (to eliminate possible start-up transients), each simulated data set was the same length as the original sample—i.e., 500 observations. Applying the six nonlinearity tests to each of these 1,000 generated data sets (and also to the sample data), we obtained the results in Table 3.

TABLE 2. GARCH(1,1) parameter estimates

Parameter	Estimate	t-value
Omega	2.157×10^{-6}	2.17
Alpha 1	0.0623	3.01
Garch 1	0.8997	25.00

TABLE 3. Summary of test results

	McLeod/ Li	Engle LM	BDS	Hinich bicovariance	Tsay	Hinich bispectrum
p-value for rejection of $H_o : x(t) \sim$ i.i.d. (sample data)	.000	.000	.083	.000	.036	.414
Estimated $E\{p_i\}$ (generated data)	.103	.120	.158	.128	.324	.392
Estimated power of 5% test[a] (generated data)	.693	.633	.484	.577	.203	.130

[a]Regarding the relationship between $E\{p_i\}$ and test power, see note 1.

The results in Table 3 indicate that the pattern of sample test results is broadly in accordance with what one might expect from the pattern of the estimated power for each of these tests: the null hypothesis that returns are identically and independently distributed (serially i.i.d.) is rejected—very strongly in three cases and at the 10% level of significance in one case—for all four of the tests that have relatively high power against this GARCH(1, 1) alternative. And the null hypothesis that returns are serially i.i.d. is not rejected for one of the two tests that have low power against this alternative. The only discrepancy is on the results for the Tsay test: here the test has relatively low power against the null hypothesis that returns are serially i.i.d., but using the sample data the Tsay test rejects this null hypothesis with p-value equal to .034.

This discrepancy was decidedly not statistically significant, however: fully 81% of the 1,000 AP test statistic values obtained using data simulated from the GARCH(1, 1) process exceeded the AP test statistic value obtained using the p-values, tabulated above, for the six tests applied to the sample data.[3] This fraction can be interpreted as the p-value at which one can reject the GARCH(1, 1) specification as adequately modeling these aspects of the process that actually generated these data.

7. CONCLUSIONS

We conclude that the GARCH(1, 1) model—which (over this sample period) is the only viable model from the ARCH/GARCH family with regard to the CRSP equally weighted index of daily returns—cannot be rejected as an appropriate model for the process generating these daily stock return data. This model thus appears to be reasonably adequate in terms of reproducing the kinds of nonlinear serial dependence addressed by this battery of nonlinearity tests, at least over the sample period (2006–2007) considered here. Patterson and Ashley (2000, Chapter 6) find strong evidence that nonlinear serial dependence in daily returns to the S&P stock market index is episodic in nature, so it is entirely possible that the

GARCH(1,1) model will be rejected in other time periods; this awaits further work.

Further, these results do not rule out the possibility that the daily returns to the CRSP equally weighted stock index are actually generated by a potentially forecastable mechanism in which current returns are a nonlinear function of past returns (plus a serially i.i.d. innovation) rather than by a GARCH process, for which only the variance of returns is forecastable. This is because nonlinearity tests are all designed to detect departures from a null hypothesis in which the return series is both independently and identically distributed. Thus, although the data simulated from the GARCH(1, 1) process (being, by construction, heteroskedastic over time) violate the "identically distributed" part of this null hypothesis, the observed rejections of this null hypothesis in the sample data may be arising instead from violations of the "independently distributed" part of this null hypothesis, which have been shown in Ashley (2009) to necessarily induce conditional heteroskedasticity in a time series. Distinguishing between these two possibilities must await further work.

NOTES

1. The estimated power of the 5% test would be the fraction of the 1,000 p-values that do not exceed .05; $E\{p_i\}$ is the average of the 1,000 p-values.

2. See Appendix 1 of Ashley and Patterson (2006) for complete descriptions and definitions of these implementing parameters. Primary references for these tests are McLeod and Li (1983), Engle (1982), Brock et al. (1986), Hinich and Patterson (2006), Tsay (1986), and Hinich (1982), respectively.

3. Essentially identical results are obtained either omitting the two nonlinearity tests explicitly focused on the detection of conditional heteroskedasticity (the McLeod/Li and Engle LM tests) or using the squared deviations from the median value rather than the squared deviations from the mean value in the AP statistic.

REFERENCES

Ashley, R. (1998) A new technique for postsample model selection and validation. *Journal of Economic Dynamics and Control* 22, 647–665.

Ashley, R. (2009) On the Origins of Conditional Heteroscedasticity in Time Series. Available at http://ashleymac.econ.vt.edu/working_papers/origins_of_conditional_heteroscedasticity.pdf.

Ashley, R. and D.M. Patterson (1986) A non-parametric, distribution-free test for serial dependence in stock returns. *Journal of Financial and Quantitative Analysis* 21, 221–227.

Ashley, R. and D.M. Patterson (2006) Evaluating the effectiveness of state-switching models for U.S. real output. *Journal of Business and Economic Statistics* 24(3), 266–277.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–327.

Bollerslev, T., R. Chou, and K. Kroner (1992) ARCH modeling in finance. *Journal of Econometrics* 52, 5–59.

Brock, W.A., W. Dechert, and J. Scheinkman (1996) A test for independence based on the correlation dimension. *Econometric Reviews* 15, 197–235.

Dalle Molle, J.W. and M.J. Hinich (1995) Trispectral analysis of stationary random time series. *Journal of the Acoustical Society of America* 97, 2963–2978.

Diebold, F.X. and R.S. Mariano (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13(3), 253–263.

Engle, Robert F. (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.

Hansen, B.E. (1999) Testing for linearity. *Journal of Economic Surveys* 13, 551–576.

Harding, D. and A. Pagan (2002) Dissecting the cycle: A methodological investigation *Journal of Monetary Economics* 49, 365–81.

Hinich, M. (1982) Testing for Gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis* 3, 169–76.

Hinich, M. and D.M. Patterson (1985) Evidence of nonlinearity in daily stock returns. *Journal of Business and Economic Statistics* 3(1), 69–77.

Hinich, M. and D.M. Patterson (2006) Detecting epochs of transient dependence in white noise. In M.T. Belongia and J.M. Binner (eds.), *Money, Measurement and Computation*, 61–75. New York: Palgrave Macmillan.

Kaplan, D.T. (1993) Exceptional events as evidence for determinism. *Physica D* 73, 38–48.

McCracken, M.W. (2007) Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics* 140(2), 719–52.

McLeod, A.I. and W.K. Li (1983) Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* 4, 269–273.

Mizrach, B. (1991) A Simple Nonparametric Test for Independence. Unpublished manuscript, Department of Economics, Rutgers University.

Nychka, D., S. Ellner, A.R. Gallant, and D. McCaffrey (1992) Finding chaos in noisy systems. *Journal of the Royal Statistical Society B* 54, 399–426.

Patterson, D.M. and R. Ashley (2000) *A Nonlinear Time Series Workshop: A Toolkit for Detecting and Identifying Nonlinear Serial Dependence*. Boston: Kluwer Academic.

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

Ramsey, J.B. and P. Rothman (1996) Time irreversibility and business cycle asymmetry. *Journal of Money, Credit, and Banking* 28, 1–21.

Saikkonen, P. and R. Luukkonen (1988) Lagrange multiplier tests for testing nonlinearities in time series models. *Scandinavian Journal of Statistics* 15, 55–68.

Tsay, R.S. (1986) Nonlinearity tests for time series. *Biometrika* 73, 461–466.

Verbrugge, R. (1997) Investigating cyclical asymmetries. *Studies in Nonlinear Dynamics and Econometrics* 2, 15–22.

White, H. (1989) Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association* 84, 1003–1013.