Using Daily Stock Returns in Event Studies and the Choice of Parametric Versus Nonparametric Test Statistics

Michael A. Berry
University of Virginia

George W. Gallinger
Arizona State University

Glenn V. Henderson, Jr.
University of Cincinnati

Abstract

This study replicates parts of the Brown and Warner [2] study of the use of daily data in event studies. A significant number of the daily returns series were nonnormally distributed. Excess returns or residuals were more normally distributed. The nonnormal nature of daily returns may suggest use of nonparametric test statistics. The current findings suggest that would be a bad choice.

Introduction

An efficient market reacts almost instantly to any relevant news event, with effects of such information reflected in security prices. This implication of the efficient market hypothesis has been tested repeatedly in event studies using variations on the residual analysis technique introduced by Fama, Fisher, Jensen, and Roll [4]. Generally, these studies find that the market does react quickly. Research has used even shorter observation intervals to determine how quickly the market reacts.

The use of short observation intervals or daily data raises methodological issues. Scholes and Williams [17] show that daily data are nonsynchronous because prices are recorded at discrete intervals, which makes accurate calculation of returns impossible. This errors in variables problem may cause regression estimates to be biased and inconsistent. Raj and Ullah [12] point out that sampling continuous variables at discrete intervals may induce nonstationary behavior in the coefficients of the structural equation being estimated. As prices move in continuous time, discrete sampling using daily data may cause stochastic coefficient behavior. Cohen *et al.* [4] explain why daily returns may behave in ways that are less than perfect for empirical research.

0747-5535/90/1600-0070 $01.00

Reinganum [13] offers several reasons for using daily data. In particular, he shows that Dimson [6], Scholes and Williams [17], and ordinary lease squares (OLS) techniques preserve rank ordering within portfolios. Using daily data increases statistical power through the additional degrees of freedom. Further, use of daily data is necessary to study daily trends or security price reactions on specific event days.

Brown and Warner (BW) often are cited in justifying the use of daily data for event studies. BW report that although daily returns are nonnormally distributed, "departures from normality are less pronounced for cross sectional mean excess returns..." [2, p. 10]. The agreement between empirical and theoretical distributions for the Student t test statistic is reassuring. (See [2, Figure 1].) Mapping the power of Student t test statistics of residuals from market model regressions is similarly satisfying [2, Figure 2].

Because of the problem of variance shifts coincident with events of interest, BW consider the use of nonparametric test statistics [2, p. 19]. They suggest that such statistics would be badly specified due to skewness in daily excess returns.

This study provides a partial replication of the BW daily study with additional consideration of the use of nonparametric test statistics. Specifically, this paper examines: the statistical characteristics of daily returns; the characteristics of residuals from a single index market model (SIMM) regression using daily data; and the specification of parametric and nonparametric test statistics when daily data are used.

The current findings, as expected, echo those reported by Brown and Warner. Daily returns are highly nonnormal. First order serial correlation in returns, as measured by a Durbin-Watson statistic, is minimal. There is significant observed heteroscedasticity. There is also significant skewness and kurtosis in the daily return series.

The prediction errors or residuals based on OLS market model regressions are more normal. Neither serial correlation nor heteroscedasticity appear to be a problem in the residual series. Using a more powerful normality test, it is concluded that the residuals are not significantly nonnormal.

The nature of the residuals distribution suggests that parametric test statistics are preferable to nonparametric ones. The empirical results support such an inference.

Methodology

This research uses simulation on a random sample of daily security returns aggregated into portfolios. On a particular day, it is posited that a hypothetical event occurs. The objective is to test how effectively the OLS methodology

detects an abnormal return on the known event date using parametric and nonparametric test statistics.

Data

A random sample of 2,000 firms was selected (with replacement) from the daily CRSP file [3] for the period 1962 through 1979. From the sample, 80 portfolios were constructed, each containing 25 randomly chosen securities. Computer capability constrained sample size.

Firms were selected according to the following procedure. First, a uniform random distribution was sampled with an arbitrarily chosen seed number. A computer algorithm transformed the random number into a CUSIP number. When a valid CUSIP was found, the first and the last trading dates were used to determine another seed for a longitudinal random number. A number, guaranteed by transformation to be between the first and last trading dates, was generated. Only firms with data for 100 consecutive trading days were chosen, with zero returns considered as valid observations. If there were 100 consecutive daily observations following the random date, that series was selected. If 100 days after were unavailable, then the 100 days before the random date were examined. If 100 days of prior observations existed, that series was selected. If neither condition was met, a period of 50 days before and after the random date was tried. If data were missing within this 100 day period, the time series was discarded. The same firm then was resampled using a different longitudinal seed. This process continued until a time series containing 100 consecutive observations was found. Of the 2,000 firms initially sampled, only nine firms did not provide 100 consecutive observations. These firms were discarded after five unsuccessful iterations. The algorithm then selected another firm using the CUSIP of the previous firm as the seed in the random selection process.

Sampling bias was minimized by assembling the portfolios as follows: for portfolios of 25 securities, security returns were placed randomly in given portfolio slots (1-80) and security slots (1-25) until all possible slots were allocated. Time series then were sorted into portfolio order and security number within portfolio. Observation showed no evident systematic bias toward a particular period or class of securities, and samples did not appear to be weighted in favor of small or large firms.

The appropriate index for event studies is also an issue. Because of the reasoning in Ohlson and Rosenberg [11] and Roll [15], this study used the CRSP value-weighted market index.

Design

Seven tests were performed on the 2,000 lognormalized return series to determine their characteristics. Autocorrelation was tested using both a Durbin-Watson (DW) test and a runs test. Variance nonconstancy was tested by dividing each return series into two equal parts and estimating the variance ratio [10, p. 136]. Three normality tests were performed; Kolmogorov-Smirnov (KS) statistics and skewness and kurtosis coefficients were examined. Finally, the mean value of each sample was tested to determine whether it was significantly greater than zero.

Residuals were calculated by running a SIMM regression on the first 90 observations of the 100 observation sample period. Abnormal performance was injected into day 91 (similar to the Brown and Warner [1, 2] studies), and residuals were calculated for days 91 through 100. Residuals series then were collected assuming alpha and beta are fixed over the ten day holdout period.

Several tests similar to those performed on the raw returns were run on the residuals to determine the robustness of the methodology to autocorrelation, heteroscedasticity, and nonnormality. Residuals were tested for serial autocorrelation, heteroscedasticity, and nonnormality. Residuals were tested for serial autocorrelation using both a DW test and a runs test. If serial correlation exists in the residual estimates, then the alpha and beta estimates are inefficient and false inferences may be drawn regarding information effects. A KS test, a Shapiro-Wilk test, and the coefficients of skewness and kurtosis were used to assess normality.

A variance ratio test was performed for each security to determine if heteroscedasticity was present in the residuals. This problem has been reported [8, 9] in the literature as causing inefficient estimates of parameters and misstatement of residual values.

Results

Return Characteristics

Table 1 presents characteristics of the 2,000 returns series. The table shows the proportion of these series that reject the respective null hypotheses at 95 percent and 99 percent significance levels. The returns were lognormalized to reduce outliers and nonnormality.

First order serial correlation, as measured by DW's D statistic, is minimal. At the 95 percent significance level, 4.9 percent of the series exhibit first order serial correlation, while 1.5 percent exhibit serial correlation at the 99 percent

level. The average D statistic of 2.027 indicates that little first order autocorrelation is present.

The runs statistics are inconsistent with the DW statistics. They indicate significant autocorrelation in 32.95 percent and 13.99 percent of the return series at the 95 percent and 99 percent levels of significance, respectively. The inconsistency between the DW and the runs results is due to the manner in which each procedure measures serial dependence. The DW measures serial correlation by testing the difference between two successive data points based on the size of that difference. The runs test is a nonparametric test and is based on the number of runs of positive and negative data points. If the differences are small, the DW test will reveal little or no serial correlation. The runs test still may find trending in the data. The runs test is a better measure of autocorrelation because it is robust to the noise in daily data.

The F statistic for heteroscedasticity is calculated using the ratio of the standard deviation of the first half of each security's returns to the standard deviation of the second half of those returns. The null hypothesis is that the variances are equal. The F statistic reveals significant variance nonconstancy in the returns data. Over 66 percent and 54 percent of the return series at the 95 percent and the 99 percent significance levels, respectively, are found to have nonconstant variance.

The KS test indicates that daily data are highly nonnormal, which is consistent with previous research [6, 17]. Normality is rejected in over 54 percent of the series at the 95 percent significance level and in almost 36 percent of the series at the 99 percent significance level.

The Student t test shows that only about 2 percent of the 2,000 series reject the null hypothesis of a mean return of zero at the 95 percent level, and less than 1 percent (about seven series) reject the mean return of zero hypothesis at the 99 percent level. In other words, the randomly selected securities in this study show no systematic abnormal returns.

The final two tests indicate that both sample skewness and kurtosis are significant. Skewness is detected in almost 54 percent of the series at the 95 percent level and in 38 percent of the sample at the 99 percent level. Kurtosis is significant in approximately 70 percent and 51 percent of the return series at the 95 percent and 99 percent significance levels, respectively. Further evidence of significant skewness and kurtosis is indicated by the average values of 3.3435 and 2.5763, respectively. These results are consistent with previous studies where both significant skewness and kurtosis were found in daily security returns.

Residual Characteristics

Table 2 reports the characteristics of the residuals conditioned with zero percent abnormal return.[1] Two general conclusions can be drawn from Table 2. First, the residuals are skewed and kurtotic. At the 99 percent significance level, skewness is present in approximately 28 percent of the residual series and kurtosis is present in approximately 31 percent of the series. At this significance level, only 1 percent of the series (20 series) should indicate significance (due to Type I error) if the residuals are normally distributed. Therefore, these results suggest that the residuals often are not normally distributed.

Autocorrelation, as measured by the DW statistic, is not a major problem. The frequency of rejection is approximately equal to the test's significance level, indicating that the residuals are well conditioned. The runs test confirms the DW statistics--again, low autocorrelation. Although the methodologies differ with respect to rejection frequencies, the differences are not statistically significant.

The F test indicates that some heteroscedasticity is present in the residuals. The average F statistic, however, indicates that heteroscedasticity is much lower in the residuals than it is in the raw return series reported in Table 1. Heteroscedasticity does not appear to be a significant statistical problem in OLS residuals.

The KS test on the residuals indicates that they are (reasonably) normally distributed. Only about 5 percent to 10 percent of the residuals are nonnormal at the 95 percent significance level, with approximately 2 percent to 4 percent of the residuals exhibiting nonnormality at the 99 percent significance level. The normality test is conducted by standardizing the residuals to determine if the distribution from which the transformed variable is generated is a standard normal distribution [10, p. 107]. The skewness and kurtosis statistics that suggested significant nonnormality therefore are inconsistent with the KS results.

Shapiro and Wilk (SW) [16] develop a test statistic that is more powerful than either the KS test or the normalized third or fourth moments in detecting nonnormality. The Shapiro-Wilk statistic, which was calculated for each residual series to provide added insight on normality of the residuals, indicates less deviation from normality, suggesting that the residuals are normally distributed. The SW statistic rejects the hypothesis of normality 4.8 percent of the time at the 95 percent significance level and .78 percent of the time at the 99

[1]Similar tests were run on residuals conditioned with 2 percent abnormal return and on series where nonsynchronous trading effects were eliminated. The characteristics of these residuals were entirely consistent with those reported in Table 2.

percent significance level. These results, in conjunction with the KS results, and the fact that the average values for the normalized third and fourth moments are not significant at the 99 percent significance level mitigate in favor of not rejecting the null hypothesis of normality for the 2,000 residual series. Therefore, it is concluded that significant nonnormality does not exist in the unadjusted residual series.

Using daily data in a SIMM OLS regression produces heteroscedastic residuals but the residuals are essentially normal. Because the residuals are heteroscedastic, power tests are necessary to test the robustness of the event methodology using daily data.

Power Tests

The most important characteristic of an event study methodology is its ability to find abnormal performance. This characteristic is the power of the methodology. Power statistics were derived for the Student t (t), the Wilcoxon signed rank (W), and the sign (S) statistics at five abnormal performance levels. The nonparametric Wilcoxon signed rank and the sign test may be better specified than the parametric Student t test because of the daily return nonnormality (shown in Table 1).

Table 3 reports the rejection frequencies (based on 80 replications) at the 99 percent, 95 percent, and 90 percent levels of significance. The amount of abnormal performance injected into the series on the hypothetical event day ranges from 0 percent to 2 percent. The Type I error rate is defined by the rejection frequencies at the 0 percent performance level. This allows inferences to be made about the correctness of the statistic. The power of methodologies is assessed by examining rejection frequencies for the various abnormal performance levels.

Type I Error

The 2,000 series of residuals in the sample were aggregated into 80 portfolios of 25 securities. (Portfolio sizes of 50 securities and 100 securities also were examined, with results similar to those reported for 25 security portfolios.) In Table 3, for the 0 percent level of abnormal performance, the Student t statistic rejects the null hypothesis at approximately the test's significance level, indicating that the test is well specified.[2]

[2]This is consistent with the findings in Brown and Warner [2, pp. 11-14].

The rejection frequencies of the nonparametric statistics produced disturbing results. The Wilcoxon and sign statistics do not behave well across significance levels. For example, at the 99 percent significance level, the Wilcoxon and the sign statistic reject the null hypothesis too seldom (actually never). The sign statistic generally produces too few rejections of the null hypothesis. These results are consistent with Brown and Warner's findings using monthly data [1].

Type II Error

Table 3 indicates that rejection frequencies increase for all statistics[3] as the normal performance level rises (with the exception of one observation that does not change, an anomaly, indicated by an asterisk, occurring at the 99 percent significance level). When abnormal performance is injected into the data, the nonparametric statistics (W and S) appear uniformly more powerful[4] than the Student t at all positive abnormal performance levels. Such tests, however, reject too few at 0 percent.

Results suggesting that the nonparametric statistics are more powerful than the Student t tests, particularly at the higher abnormal performance levels, are unexpected. Brown and Warner [1], using monthly returns and an equal-weighted index, find that nonparametric statistics reject too seldom when abnormal performance is present. A possible explanation for this inconsistency is that the nonparametric statistics are relatively more powerful with daily data; that is, with more degrees of freedom. That contradicts the earlier finding that the nonparametric statistics fail to reject the null hypothesis often enough when no abnormal performance is present. The point is crucial; an analyst may prefer to use nonparametric statistics, ones that appear more powerful at higher levels of abnormal performance. Nonparametric tests, however, do not appear well specified under the null hypothesis; Type I error is too high.

Sampling Distributions

Given the apparent greater power of the nonparametric tests (and some contradictory results), an examination of the characteristics of the sampling distributions for each of the statistics is in order. The nonparametric tests

---

[3]The calibration of these statistics (determining the minimum amount of abnormal performance that each can detect) provides information enabling researchers to determine if the event study framework is suitable for a given event based on the expected abnormal return.

[4]It should be noted that several different conclusions can be reached concerning such a characteristic, depending upon a researcher's preferences concerning Type I error.

assume that 50 percent of the observations are positive and 50 percent are negative; given the skewness in daily returns, there is reason to believe that these statistics are misspecified [1]. The results shown in Table 4 are based on portfolios of 25 securities. Again, portfolios of 50 securities and 100 securities also were analyzed, with results that do not change materially as portfolio size is increased.

Table 4 presents the characteristics of the sampling distributions of the Student t statistic, the Wilcoxon signed rank statistic, and the sign statistic. These results are derived from the residuals conditioned with 0 percent abnormal performance.

The three test statistics should be distributed as follows if they are well specified: the Student t statistic should be distributed according to the Student distribution, while the Wilcoxon and sign statistics should be distributed according to the normal density function with mean zero and standard deviation equal to one. Table 4 indicates that the Student t statistic is well specified. The nonparametric Wilcoxon and sign statistics are misspecified consistently. For example, the probability of observing a KS D statistic as large as .0876 (the KS statistic that corresponds to the sampling distribution for the Student t test) is 57 percent; therefore, it can be inferred that the t test is well specified. Conversely, both the Wilcoxon signed rank and sign statistics are poorly specified. There is less than a 1 percent probability (.0053) of observing a KS statistic of .1926 of the Wilcoxon statistic is truly N(0,1) as the test assumes. The results are similar (worse) for the nonparametric sign test.

These results show that the Student t statistic is an appropriate test statistic for event studies with daily (nonnormal) data. This is consistent with the Brown and Warner [1] results using monthly data. Even though nonparametric statistics may have greater power, researchers should exercise caution in drawing inferences from their use as they understate the probable Type I error.

Summary and Conclusions

Daily return data create problems for event studies that use residual analysis. Although autocorrelation does not appear to be present, the return series exhibits several other disturbing characteristics; that is, returns generally are not normal. They show significant skewness and kurtosis. Further, a significant proportion of the returns are heteroscedastic.

The residuals from a typical OLS SIMM regression of these returns, however, are much better conditioned. Some minor heteroscedasticity remains, but probably not enough to cause problems. More refined tests suggest that the residuals from the regression are not significantly nonnormally distributed.

The statistical power results are informative. Although the nonparametric statistics appear to be more powerful at detecting abnormal performance, their Type I error characteristics and their sampling distribution leads to the conclusion that they are ill specified and should be used only with extreme care.

To summarize, daily data leave something to be desired with regard to their statistical properties for use in OLS regressions. At the same time, the residuals from such regressions are surprisingly well conditioned for statistical testing. The Student t works well with such residuals, Type I error is appropriate, and the sampling distribution is well specified. The same cannot be said for the nonparametric statistics. Their apparent power for finding abnormal returns is attractive, but their ill-specified sampling distributions should put the potential user on guard. The Student t seems much less treacherous. Brown and Warner [2] recommend it. This research also recommends it and specifically warns against using the nonparametric alternatives.

References

1. Brown, S. and J. Warner, "Measuring Security Price Performance," *Journal of Financial Economics* (September 1980), pp. 205-258.

2. Brown, S. and J. Warner, "Using Daily Stock Returns: The Case of Event Studies," *Journal of Financial Economics* (March 1985), pp. 3-31.

3. CRSP Daily Returns File, Center for Research in Security Prices, University of Chicago (1981).

4. Cohen, K.J., G.A. Hawawini, S.F. Maier, R.A. Swartz, and D. K. Whitcomb, "Implications of Microstructure Theory for Empirical Research on Stock Price Behavior," *Journal of Finance* (May 1980), pp. 249-257.

5. Dowen, R.J. and S.C. Isberg, "Reexamination of the Intervaling Effect on the CAPM Using a Residual Return Approach," *Quarterly Journal of Business and Economics* (Summer 1988), pp. 114-129.

6. Dimson, E., "Risk Measurement When Shares are Subject to Infrequent Trading," *Journal of Financial Economics* (June 1979), pp. 197-226.

7. Fama, E., L. Fisher, M. Jensen, and R. Roll, "The Adjustment of Stock Prices to New Information," *International Economic Review* (February 1969), pp. 1-21.

8. Martin, J. and R. Klemkosky, "Evidence of Heteroscedasticity in the Market Model," *Journal of Business* (January 1975), pp. 81-86.

9. Miller, J. and M. Scholes, "Rates of Return in Relation to Risk: A Re-Examination of Some Recent Findings," in M. Jensen (ed.), *Studies in the Theory of Capital Markets* (New York: Praeger, 1972).

10. Neter, J. and W. Wasserman, *Applied Linear Statistical Models* (Homewood, IL: Richard D. Irwin, Inc., 1974).

11. Ohlson, J. and B. Rosenberg, "Systematic Risk of the CRSP Equal-Weighted Common Stock Index: A History Examined by Stochastic-Parameter Regression," *Journal of Business* (January 1982), pp. 121-145.

12. Raj, B. and A. Ullah, *Econometrics, A Varying Coefficient Approach* (New York: St. Martins Press, 1981).

13. Reinganum, M., "A New Empirical Test of the CAPM," *Journal of Financial and Quantitative Analysis* (November 1981), pp. 439-462.

14. Roll, R., "Ambiguity When Performance is Measured by the Securities Market Line," *Journal of Finance* (September 1978), pp. 1051-1069.

15. _____, "A Possible Explanation of the Small Firm Effect," *Journal of Finance* (September 1981), pp. 879-888.

16. Shapiro, S. and M. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika* (December 1965), pp. 591-611.

17. Scholes, M. and J. Williams, "Estimating Betas from Nonsynchronous Data," *Journal of Financial Economics* (December 1977), pp. 309-327.

Table 1
Statistical Characteristics of Raw Returns
(2,000 observations)

Sig. Level	Durbin Watson (D)	Runs (Z)	Heter (F)	Normality (KS)	T (t)	Skewness B(1)	Kurtosis B(2)
.95	0.049	0.3295	0.6615	0.5435	0.0235	0.5360	0.6995
.99	0.015	0.1399	0.5450	0.3580	0.0035	0.3835	0.5140
Ave.	2.027	-1.4407	2.4677	1.4920	0.1197	3.3435	2.5763
				Critical Values for Each Statistic			
.95	1.500	1.960	2.270	1.960	2.000	0.533	0.990
.99	1.310	2.570	2.970	2.580	2.670	0.787	1.880

Table 2
Statistical Characteristics of Residuals
(2,000 observations)

Sig. Level	Durbin Watson (D)	Runs (Z)	Heter (F)	Normality (KS)	T (t)	Skewness B(1)	Kurtosis B(2)
.95	0.0720	0.0890	0.2490	0.0760	0.0190	0.4380	0.4890
.99	0.0170	0.0195	0.1390	0.0160	0.0010	0.2830	0.3050
Ave.	2.049	0.271	1.290	0.872	-0.0410	0.286	1.831
				Critical Values for Each Statistic			
.95	1.500	1.960	2.270	1.960	2.000	0.533	0.990
.99	1.310	2.570	2.970	2.580	2.670	0.787	1.880

Table 3
Power Tests at Various Significance Levels
(proportion of rejected null hypotheses)

Test	0%	.02%	Injected Performance .5%	1%	2%
			99% Significance Level		
t	0.025	0.038	0.063	0.125	0.762
W	0.000	0.038	0.100	0.425	0.950
S	0.000	0.000*	0.050	0.313	0.850
			95% Significance Level		
t	0.063	0.075	0.125	0.388	0.950
W	0.050	0.125	0.313	0.600	0.988
S	0.013	0.075	0.225	0.550	0.963
			90% Significance Level		
t	0.075	0.100	0.263	0.600	0.988
W	0.088	0.213	0.388	0.738	1.000
S	0.050	0.175	0.325	0.750	0.988

*One anomaly where rejection frequency did not increase with the level of injected abnormal return.

Table 4
Characteristics of the Sampling Distributions

Test Statistic	Mean	Var	Skew	Kurt	KS	PR>Z
t	-0.0602	0.9535	0.4011	3.8317	0.0876	0.5713
W	-0.8820	0.3469	0.5374	3.6374	0.1926	0.0053
S	-1.7500	888.5443	0.0095	1.0048	0.5250	0.0000