Journal of Empirical Finance 16 (2009) 466–482

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Sample selection and event study estimation☆

Kenneth R. Ahern*

Ross School of Business, University of Michigan, 701 Tappan Street, Ann Arbor, MI 48109, United States

ARTICLE INFO

Article history:
Received 23 April 2008
Received in revised form 15 January 2009
Accepted 23 January 2009
Available online 31 January 2009

JEL classification:
G30
C14
C15

Keywords:
Event studies
Nonparametric test statistics
Multifactor models
Characteristic-based benchmark model

ABSTRACT

The anomalies literature suggests that pricing is biased systematically for securities grouped by certain characteristics. If these characteristics are related to selection in an event study sample, imprecise predictions of an event study method may produce erroneous results. This paper performs simulations to compare a battery of short-run event study prediction and testing methods where samples are grouped by market equity, prior returns, book-to-market, and earnings-to-price ratios. Significant statistical errors are reported for both standard and newer methods, including three- and four-factor models. A characteristic-based benchmark model produces the least biased returns with the least rejection errors in all samples.

1. Introduction

The wide variety of applications and the richness of data available have made event studies commonplace in economic, finance, and accounting research. The strength of the event study methodology is that abnormal returns due to a firm-specific, but time-independent event may be precisely estimated by aggregating results over many firms experiencing a similar event at different times. Brown and Warner (1985) (BW) conduct simulated event studies of random samples and find that simple estimation techniques are well specified. In particular, estimates from ordinary least squares (OLS) with a market index tested with parametric statistical tests are well-specified using non-normally distributed daily data and in the presence of non-synchronous trading. Moreover, BW shows that abnormal returns measured with simpler estimation procedures such as market-adjusted and mean-adjusted returns display no significant mean bias.

This paper suggests that the results of BW may not hold in actual event studies. BW show results for data that are randomly selected from all securities, whereas event studies typically have data that are characteristically non-representative of the overall market and often grouped by underlying traits such as size, momentum, and valuation. For instance, firms that initiate dividends, split their stock, or make acquisitions are likely to be large with high prior returns. Under these conditions, it should not be assumed that the market average results of BW should hold.

This paper simulates event studies similar to BW, but draw samples non-randomly. In particular, samples are drawn from the highest and lowest deciles of market equity, prior returns, book-to-market, and earnings-to-price ratios, where deciles are computed from all NYSE firms. I run a horse race between eight prediction methods, including a characteristic-based benchmark model, a market model, Fama French Three-Factor and Carhart Four-Factor models, and four test statistics, both parametric and

☆ I thank Richard Roll for encouraging me to develop this topic. I especially appreciate the comments of Stephen Brown and Jerold Warner. I also thank Jean-Laurent Rosenthal, Antonio Bernardo, J. Fred Weston, Duke Bristow, and Raffaella Giacomini for helpful comments. The comments of the associate editor and two anonymous referees substantially improved this paper. This paper is an extension of Chapter 1 of my dissertation completed at the Economics Department of UCLA.

* Tel.: +1 734 764 3196.
E-mail address: kenahern@umich.edu.

0927-5398/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jempfin.2009.01.003

non-parametric, to determine which method has the least mean bias and the best power and specification of the tests in different samples. I also investigate the effect of using post-event versus pre-event data to estimate model parameters, the effect of event-induced variance, and the relation between bias and sample size.

The results suggest that standard event study methods produce statistical biases in grouped samples. The most significant errors are found to be false positive abnormal returns in samples characterized by small firms and firms with low prior returns. False negative abnormal returns are found in samples characterized by large firms and by firms with high prior returns. The characteristic-based benchmark model, where stock returns are adjusted by a matched size-return portfolio of control stocks, displays the least bias of all the models. Multifactor models produce only marginal benefits over a standard market model in predicting event day normal returns, but they generate less skewed abnormal returns that are better suited for statistical tests. Using post-event estimation windows also reduces forecast error bias. Event window variance increases present a bigger problem than pricing bias, but may be corrected using the sign test.

Though the mean bias reported in the prediction models is small, it significantly affects the Type I and Type II errors of the models. With the exception of the characteristic-based benchmark model, all of the models over-reject the null hypothesis of no abnormal returns in lower tail tests in samples of large firms. Likewise, they under-reject in upper tail tests for large firms. In samples grouped by prior returns, the models over-reject in upper tail tests in samples of firms with low prior returns and in lower tail tests of samples of firms with high prior returns. The characteristic-based benchmark model also suffers from this problem, but to a lesser degree than the standard methods. The bias in rejection rates across tails affects the asymmetry of the power of the tests in similar ways.

Samples composed of multivariate distributions across the mis-pricing characteristics may be prone to compound biases in some cases, or may have reduced errors if the pricing biases cancel out. Running simulations where samples are matched to distributions of size and prior returns from 564 stocks that completed reverse-stock splits during 1972 to 2002, I find a slight mean bias in all models, but the economic significance is small. In addition, other simulations are performed with samples intended to match large acquirers, distressed firms, firms with new exchange listings, and firms making seasoned equity offerings. Biases are not large in general, but when actual event returns are small or zero (such as in acquisitions), or if the event window is long to account for noisy measures of event dates, the biases will dominate the actual returns. In addition, the return biases lead to large abnormal dollar return biases in samples of large firms ($45 million over three days for a firm at the 85th NYSE percentile of market equity).

Other studies have addressed sample selection bias. Brown et al. (1995) present a theoretical model of survivorship where volatility is positively related to an ex post bias produced by a lower bound on prices for surviving firms. Thus, this bias would be severe for firms in financial distress, for example, which may be an explicit inclusion condition in an event study of corporate restructuring. Brown et al. also show that a sample of stock splits will be conditioned on the occurrence of positive returns in the pre-event period. Bias introduced by a size effect is analyzed in Dimson and Marsh (1986) for the case of press recommendations on stock returns. They also note that stock price run-ups may attract the attention of the press and lead to more recommendations for firms with high prior returns. Fama (1998) states that even risk adjustment with the true asset pricing model can produce sample specific anomalies if sample specific patterns are present. These studies suggest that the underlying characteristics of firms selected for an event study sample may lead to biased predictions if a non-robust prediction technique suitable for market average firms is used.

The present paper provides a number of contributions to the event study methods literature. Though others have recognized that firm characteristics associated with pricing anomalies may be correlated with corporate events, there has not been a comprehensive simulation study of grouped samples to determine if the potential biases are small enough to be ignored. Second, the random sample results of BW will be significantly updated. The data in Brown and Warner cover the seventeen years from 1963 to 1979, of which only seven include NASDAQ firms (1973–1979). The time period in my study extends the data of BW to almost 40 years (1965–2003), with 31 years including the NASDAQ. This will provide a much broader universe of securities for which to test the specification of event study methods. Finally, this is the first paper to study the benefits of using multifactor and characteristic-based benchmark models in short-run event studies. Combining a much larger dataset with a comprehensive collection of prediction models and test statistics in an event study simulation using daily returns data will bring evidence to bear on the potential biases in event study methodologies. It should be noted that the appropriate choice of methodology will depend upon the setting of a study and that the results of this paper do not provide a 'best' method suited to all studies.

2. Methodological issues of event studies

2.1. Selection by security characteristics

If event study sample securities are characterized by factors related to pricing biases, then the abnormal returns estimated by the event study are potentially biased. These prediction biases arise due to skewness in returns, even after including an intercept term. Take for example the well known small-firm bias in the CAPM (Banz, 1981). Suppose the true population model is:

$$R_i = \alpha + \beta_1 R_i^{\mathrm{m}} + \beta_2 \mathrm{SMB}_i + \varepsilon_i, \tag{1}$$

where R_i is an individual firm return, R^{m} is the market return, and SMB is the size factor in Fama and French (1993). However, instead of the true model we estimate:

$$R_i = \alpha + \beta_1 R_i^{\mathrm{m}} + \varepsilon. \tag{2}$$

Following Wooldridge (2000), the estimated coefficients of β_1 and α are

$$\hat{\beta}_1 = \beta_1 \mathrm{SST}_1 + \beta_2 \sum_{i=1}^{n} \left(R_i^{\mathrm{m}} - \overline{R^{\mathrm{m}}}\right)\mathrm{SMB}_i + \sum_{i=1}^{n} \left(R_i^{\mathrm{m}} - \overline{R^{\mathrm{m}}}\right)\varepsilon_i \tag{3}$$

$$\hat{\alpha} = \overline{R_i} - \hat{\beta}_1 \overline{R^{\mathrm{m}}} \tag{4}$$

where SST_1 is the sum of squared deviations in R_i^{m}. The omitted variable bias will lead to the following incorrect estimates of α and β_1:

$$E\left(\hat{\beta}_1\right) = \beta_1 + \beta_2 \hat{\delta} \tag{5}$$

$$E\left(\hat{\alpha}\right) = \overline{R_i} - \hat{\beta}_1 R_{\mathrm{m}} \tag{6}$$

where $\hat{\delta}$is the estimated coefficient from a regression of SMB on R_{m} and a constant. However, even with these coefficient biases, the estimate of the average firm return is unbiased,

$$E(R_i) = \overline{R_i} \tag{7}$$

Since the slope coefficient is biased, the intercept term adjusts so that the fitted line predicts the mean firm return given the mean market return. If on any given event day we expect to observe the average market return, then this fitted line would correctly predict the firm return and no prediction error would occur. Moreover, prediction errors that occur when the market return is higher or lower than its average would cancel each other out.

However, it is incorrect to expect to observe the average market return on any given day because returns are positively skewed. This implies that the mean is larger than the mode in the distribution and that for more than half of a sample of randomly chosen days, the observed return will be less than the mean return. Because of the skewness, omitted variable bias will generate incorrect predictions on average even when the model allows for an endogenously determined intercept term. This bias will persist even in large samples. Thus, because biases in standard asset pricing models are generated by omitted variables, it makes sense to look at samples grouped by these characteristics.

As mentioned, Banz (1981) finds that the CAPM predicts returns that are too low for small firms. Basu (1983) shows that price to earnings is negatively related to returns, controlling for market beta, suggesting that the CAPM will predict returns too high for firms with high P/E ratios. It is hypothesized that a simple market model event study procedure will make the same mistakes, leading to a false finding of positive abnormal returns for small firms and negative abnormal returns for firms with high P/E ratios.

Pricing anomalies due to momentum also have been documented. Jegadeesh and Titman (1993) find that securities exhibiting recent (past year) levels of high (low) returns have predictably high (low) returns in the following three months after accounting for systematic risk and delayed reactions to common factors. Thus short-run pricing models under-price securities with high returns in the recent past, and over-price securities with low returns. This may lead to the appearance of positive abnormal returns to positive momentum firms and negative abnormal returns to negative or low momentum firms. If returns are mean-reverting, however, the bias would have the opposite sign.

Book-to-market ratios also have been found to predict returns systematically. Fama and French (1992) find a positive relation between average return and book-to-market equity ratios after accounting for beta. This suggests event studies using a prediction model with only a market index as an explanatory variable will tend to find false negative abnormal returns for firms with low book-to-market ratios and false positive abnormal returns for firms with high book-to-market ratios.

These pricing anomalies may confound event study results if samples are dominated by securities characterized by the above factors. Prior studies have shown that firms that undergo particular corporate events often have common characteristics of size, momentum, and valuation ratios different from market averages. Table 1 presents a summary of the sample characteristics of prior event studies.[1] Samples of large firms with high prior returns and low book-to-market ratios are typical in studies of acquisitions and stock splits. Both samples of new exchange listings and acquisition targets tend to have small firms with high prior returns and low book-to-market ratios. Seasoned-equity offerings, dividend initiations and omissions, bankruptcy, and other corporate events also have samples that differ from market averages across these pricing anomaly factors. Given the non-random samples of prior event studies, it is relevant to determine if abnormal returns generated by standard event study methods are systematically biased when samples are grouped by the above characteristics.

2.2. Prediction and testing methods

This paper performs Brown and Warner event study simulations when samples are grouped by characteristics associated with the possible inclusion into an actual event study, namely market equity (ME), prior returns (PR), book-to-market (BM), and

[1] Earnings-to-price ratios are not reported in Table 1 because they are typically not reported in event studies. However, it is reasonable to assume that event firms have non-random E/P ratios given the preponderance of non-random samples across size, momentum, and book-to-market ratios. Thus E/P is included in this study as a potential source of bias.

Table 1
Sample characteristics across pricing anomalies.

Event	Sample firm characteristics			
	Market equity	Prior returns	Book-to-market	Source
Acquisition	High	High	Low	Rhodes-Kropf et al. (2005)
	High	N/A	Low	Mitchell and Stafford (2000)
	N/A	High	N/A	Asquith (1983)
Bankruptcy	Low	Low	—	Campbell et al. (2005)
Dividend initiation	High	High	—	Lipson et al. (1998)
	High	High	N/A	Michaely et al. (1995)
Dividend omission	High	Low	N/A	Michaely et al. (1995)
New exchange listing	Low	High	Low	Dharan and Ikenberry (1995)
Seasoned equity offering	Low	High	Low	Brav et al. (2000)
	Low	N/A	Low	Mitchell and Stafford (2000)
Share repurchase	Low	N/A	Low	Mitchell and Stafford (2000)
Stock split	High	High	Low	Ikenberry et al. (1996)
Target of acquisition	Low	N/A	Low	Rhodes-Kropf et al. (2005)
	Low	High	Low	Schwert (2000)
	N/A	Low	N/A	Asquith (1983)

This table presents the relationship between sample firms in prior event studies to market averages for the characteristics of market equity, prior returns, and book-to-market ratios. High (Low) indicates sample firms are above (below) market averages for a particular characteristic. No difference between the market average and the sample average is indicated by —. N/A indicates that the information was not reported in the source article.

earnings-to-price ratios (EP). For each sample criterion, the properties of a battery of prediction models and test statistics will be examined.

The prediction models tested in this study include traditional methods as well as less commonly used methods for comparison. The simplest method used to predict a normal return is to simply subtract a security's time series average from an event date return (mean-adjusted return), denoted here as MEAN. The most commonly used prediction method is the market model, where firm returns are regressed on a constant term and a market index, either equal- or value-weighted (MMEW and MMVW). A similar procedure, the market-adjusted return method, subtracts the market index from an event date security return (MAEW and MAVW). In both the market model and the market-adjusted return procedures researchers need to choose a market index. Because the criteria for this choice are not well defined, this paper analyzes both equal- and value-weighted indexes for comparison.

In response to the pricing anomalies of the CAPM discussed above, alternative pricing models have been developed, though their use in short-run event studies has been limited. In particular, Fama and French (1996) use a three-factor model including a market index, size index, and book-to-market index to explain stock returns (FF3F). Carhart (1997) uses a four-factor model which appends the Fama–French three-factor model with a short-run momentum index (FF4F). Both of these models will be tested alongside the more common prediction models.

Though multifactor regression models may alleviate the omitted variable bias of a simple market model, they may also introduce additional estimation error (Fama and French, 1997). An alternative to regression models is a characteristic-based benchmark (CBBM) estimate as used in the mutual fund performance literature (Daniel et al., 1997). In this model, the event returns of a sample stock are adjusted by the equally-weighted returns of a portfolio of ten control stocks matched by size and prior-return deciles. This model has the advantage that no regressors need to be estimated, which reduces estimation error. In addition, this model does not require a researcher to choose a 'normal' estimation period, either pre- or post-event. The details of all models are provided in Appendix A.

The four leading test statistics used in event studies, *t*-statistic, standardized *t*-statistic, rank, and the sign statistics, are compared in this study. The *t*-statistic is computed as in Brown and Warner (1985). The three remaining statistics are calculated as described in Corrado and Zivney (1992). The standardized *t*-statistic normalizes each abnormal return by the firm's time series standard deviation. The rank test orders the abnormal returns over the entire period that includes both the estimation and event windows and assigns a corresponding value between zero and one for each day's observation. The sign test assigns either a negative one, a positive one, or zero to each day's observation for abnormal returns that are above, below, or equal to the median abnormal return, respectively. Thus, if an event date has an average ranked abnormal return across firms close to one, or a majority of abnormal returns above the median abnormal return, then the rank and sign tests will reject the null hypothesis of no abnormal returns. Because the non-parametric tests are based on medians and do not require distributional assumptions, they may be more appropriate for the skewed and highly kurtotic nature of daily stock returns. Details of all test statistics are provided in Appendix A.

In response to the problems of pricing anomalies in event studies, researchers have used estimation periods other than the period immediately prior to the event period, though this has typically been done in limited circumstances and generally for long-run studies using monthly data.[2] Mandelker (1974) addresses this issue by separately estimating parameter coefficients using both

[2] I am grateful to Jerold Warner for suggesting this approach.

pre- and post-event estimation period data on mergers. Copeland and Mayers (1982) use post-event data in order to minimize bias associated with abnormal prior returns in the pre-event period for firms ranked by Value Line. Agrawal et al. (1992) and Gregory (1997) use post-event estimation data in long-run studies of mergers. The present study estimates all models with separate pre- and post-event estimation windows. Unless otherwise noted, all results presented in this paper will be generated using pre-event data, since this is the more common procedure.

Following BW and other simulation studies, this paper artificially introduces abnormal returns as well as variance increases to the event date returns for each characteristic sample. This facilitates comparisons between the prediction model-test statistic combinations to determine which methods are the best specified and have the most power to detect abnormal returns.

This study will concentrate only on short-run event study methods, restricting analysis to a one-day event window. This provides the best comparison of the various methods because the shorter the event window, the more precise are the tests. If a test does not perform well for a one-day event window, it will only perform worse for longer-run studies. Thus if small errors are presented in this study, they will be compounded in long-run studies (Fama, 1998; Kothari and Warner, 2005). Moreover, recognizing the problem of predicting normal returns over a long horizon, long-run event studies use different methodologies than those presented here.[3]

3. Experimental design

This study simulates 1000 samples of 250 securities each by random selection with replacement from a subset of securities in the CRSP Daily Stock dataset between January 1965 and December 2003, where subsets are based on size, momentum, and two measures of valuation. Abnormal returns are generated and tested by the introduction of artificial performance and variance on event date returns. Each of these topics is discussed in detail below.

3.1. Data requirements

To be included in this study a security must meet the following requirements. It must be an ordinary common share of a domestic or foreign company (CRSP SHRCD = 10, 11, or 12). This excludes ADRs, SBIs, closed-end funds, and REITs. Furthermore, it must not be suspended or halted (CRSP EXCHCD = 0, 1, 2, or 3). For each security-event date, (day 0), the daily returns are collected over a maximum period of 489 days (−244, +244) where the pre-event estimation period is defined as (−244, −6), the event period is (−5, +5), and the post-event estimation period is (+6, +244). However, if a firm has at least 50 non-missing returns in the pre-event estimation period, at least 50 non-missing returns in the post-event estimation period, and no missing observations in the period (−15, +15), then it is included in the sample. If an observation is CRSP coded -99, the current price is valid, but the previous price is missing, which means that the next return is over a period longer than one day, so observations following a -99 code are counted as missing observations.

Event dates are randomly chosen over all trading days between January 1, 1965 and December 31, 2003, so that each month and day of the year is equally represented in the sample. For a randomly selected event date, a security is randomly chosen from the grouped samples. If this security does not meet the requirements listed above, a new security is selected using the same event date. If no security in the sample meets the requirements for inclusion on a particular event date, a new event date is chosen. This is done to ensure that event dates are evenly distributed over the 39 years, even though there are many more possible security-event dates in later years due to a greater number of firms in the dataset. Since the focus of this paper is mean bias, event-date clustering is not investigated. Because cross-sectional dependence of samples grouped by size, prior returns, or value is larger than random samples, biases in standard errors using OLS in clustered event dates will be larger in the grouped samples. However, using a portfolio approach or seemingly-unrelated-regressions will correct the bias, just as in random samples (Campbell et al., 1997).

3.2. Sample characteristics

For each security that meets the share type and active trading requirements listed above, samples are formed on the characteristics of market capitalization (ME), prior returns (PR), book-to-market (BM), and earnings-to-price ratios (EP). These measures are calculated as in Fama and French (1992) and Kenneth French's Web site.[4] Each security is then assigned a quadruple decile according to the New York Stock Exchange (NYSE) breakpoints provided on Kenneth French's Web site for the four characteristics of ME, PR, BM, and EP, where the BM and EP deciles are assigned yearly and the ME and PR deciles are assigned monthly. For each yearly decile assignment, the corresponding twelve months are assigned the same decile. Thus each security is assigned a decile for each of the four characteristics for each month where data are available. If the accounting or returns data do not allow one or more of the characteristics to be computed and assigned to a NYSE decile, the security is still eligible for inclusion in a sample, though it will not be included in a sample grouped by a missing decile assignment.

[3] See Mitchell and Stafford (2000) for a discussion of the various issues that arise in ling-run event studies. Also see Barber and Lyon (1996), Kothari and Warner (1997), Fama (1998), Lyon, Barber, and Tsai (1999), Brav (2000), Loughran and Ritter (2000), Alderson and Betker (2006), and Gur-Gershgoren, Hughson, and Zender (2008) for further research on long-run event study methodologies.

[4] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Table 2
Sample return properties over estimation period.

Sample		Mean	Median	Standard deviation	*t*-stat for mean	Skewness	Kurtosis	Studentized range
All CRSP		0.0008	−0.0002	0.0338	0.384	0.629	9.858	8.239
ME	High	0.0008	0.0001	0.0179	0.664	0.158	5.791	7.305
	Low	0.0008	−0.0002	0.0413	0.314	0.832	11.769	8.604
PR	High	0.0024	−0.0001	0.0380	0.995	0.897	10.391	8.027
	Low	−0.0010	−0.0007	0.0475	−0.319	0.451	11.020	8.714
BM	High	0.0014	−0.0001	0.0411	0.538	0.917	12.514	8.677
	Low	0.0009	−0.0004	0.0392	0.344	0.564	8.956	8.175
EP	High	0.0010	−0.0001	0.0313	0.515	0.740	10.369	8.405
	Low	0.0010	−0.0004	0.0446	0.365	0.770	10.218	8.341

Percentage points				
250 observations from a normal population				
Variable	0.01	0.05	0.95	0.99
t-statistic	−2.596	−1.970	1.970	2.596
Skewness	−0.360	−0.251	0.251	0.360
Kurtosis	2.42	1.55	3.52	3.87
Studentized range	4.544	4.812	6.624	7.140

Sources: Pearson and Hartley (1966, Tables 34B and C), Lund and Lund (1983).
Properties of daily event study performance measures where samples are randomly drawn from either the entire CRSP database, or are selected based on one of the characteristics of market equity (ME), one-year prior returns (PR), book-to-market ratio (BM), or earnings-to-price ratio (EP). The randomly selected event dates cover the period 1965–2003. The performance measures are computed over the estimation period (−244, −6). Measures are taken from 1000 samples of 250 securities. For each parameter, the table reports the mean of 250,000 estimates in decimal format. High indicates samples drawn from the top NYSE decile. Low indicates samples drawn from the bottom NYSE decile. The percentage points listed at the bottom describe the critical values of each statistical measure for a sample of 250 standard normal random variables, corresponding to the days of the estimation period. For example, a kurtosis of 3.52 occurs randomly 5% of the time in such a sample.

Characteristic samples are chosen by selecting securities that are assigned to a particular decile or group of deciles for a particular characteristic. Thus for a given randomly selected event date, a sample firm is selected randomly from all firms that meet the decile requirement for a particular characteristic for the previous month-end.

4. Results

4.1. Estimation period returns

Table 2 displays the distribution properties of sample returns over the pre-event estimation period (−244, −6) by characteristic samples. For each characteristic, two sample groupings are formed. *High* indicates a sample formed by only including firms ranked in the top NYSE decile for a particular characteristic. *Low* samples are formed from the bottom NYSE decile. Each number reported is the mean performance measure over the estimation period of the 250,000 firm-date sample observations.

As is well documented, random sample returns are non-normally distributed with positive skewness, leptokurtosis, and a studentized range larger than normal. Average daily returns are 0.08%. The raw returns data presented here from 1965 to 2003 have a higher mean, standard deviation, kurtosis, and studentized range, and less skewness than the earlier period returns over 1963–1979, as presented in BW. This is not surprising, since the present data include many more listings of small firms.

Mean returns are identical between large and small firms and equal the random sample mean returns of 0.08%. However, high ME firms have returns with smaller standard deviations, skewness, kurtosis, and studentized ranges than the random sample firms. High prior return firms have a time-series average return of 0.24%, three times as large as random samples. Low prior return firms have negative daily returns of −0.1% on average. High PR firms also have smaller standard deviations, kurtosis, and studentized ranges than low PR firms, but greater positive skewness.

High BM firms are characterized by performance measures above the random sample benchmarks, including high mean returns, standard deviations, skewness, kurtosis, and studentized range. The low BM firms exhibit returns very similar to random sample returns. Both the highest and lowest EP deciles have returns with means above the random sample mean, though there is a negative relationship between EP and standard deviation. The other statistical measures of skewness, kurtosis, and studentized range are quite similar between the two deciles and are above the values of the random sample measures. This suggests that the performance measures of EP grouped firm returns display non-linear patterns across NYSE deciles.

4.2. Prediction model performance on day 0

Table 3 presents cross-sectional results for returns and prediction model abnormal returns on day '0.' These values reflect the ability of the prediction model to accurately predict a 'normal' return in the event period. Panel A of Table 3 presents the

Table 3
Mean performance measures at day 0.

Model	Random sample	ME		PR		BM		EP	
		High	Low	High	Low	High	Low	High	Low
Panel A: Models independent of estimation period									
Returns	0.0010***	0.0004***	0.0013***	0.0008***	0.0012***	0.0016***	0.0007***	0.0012***	0.0011***
CBBM	0.0001	−0.0000	0.0001	−0.0001	−0.0001	−0.0000	0.0001	0.0000	0.0001
MAEW	0.0002**	−0.0004***	0.0004***	−0.0001	0.0003**	0.0007***	−0.0002*	0.0003***	0.0003***
MAVW	0.0006***	−0.0000	0.0009***	0.0004***	0.0007***	0.0011***	0.0003***	0.0007***	0.0007***
Panel B: Abnormal returns generated using the pre-event estimation period									
MEAN	0.0002**	−0.0003***	0.0005***	−0.0016***	0.0021***	0.0002	−0.0001	0.0002*	0.0001
MMEW	0.0002**	−0.0004***	0.0005***	0.0016***	0.0020***	0.0002	−0.0002**	0.0001	0.0001
MMVW	0.0002**	−0.0003***	0.0004***	0.0016***	0.0020***	0.0001	−0.0002*	0.0002*	0.0001
FF3F	0.0002*	−0.0003***	0.0004***	0.0016***	0.0019***	0.0001	−0.0002**	0.0001	0.0001
FF4F	0.0002*	−0.0003***	0.0004***	0.0016***	0.0019***	0.0001	−0.0002**	0.0001	0.0002
Panel C: Abnormal returns generated using the post-event estimation period									
MEAN	0.0001	−0.0001	0.0000	0.0001	−0.0004***	0.0003	0.0000	0.0001	0.0000
MMEW	0.0001	0.0000	0.0000	−0.0000	−0.0004***	−0.0000	0.0000	0.0001	0.0000
MMVW	0.0001	−0.0000	0.0000	0.0000	−0.0004***	−0.0001	0.0000	0.0001	−0.0000
FF3F	0.0001	−0.0000	0.0001	0.0000	−0.0003**	−0.0000	0.0001	0.0001	0.0000
FF4F	0.0001	−0.0000	0.0001	0.0000	−0.0003**	−0.0000	0.0000	0.0001	0.0000

Cross-sectional properties on day 0 over samples randomly drawn from either the entire CRSP database or the highest or lowest NYSE decile of each characteristic market equity (ME), one-year prior returns (PR), book-to-market ratio (BM), or earnings-to-price ratio (EP). The randomly selected event dates cover the period 1965–2003. Models are described in the appendix. Each number reported is based on 1000 values of sample mean measures, where the sample mean measure is the average (adjusted) return over the 250 securities in each sample. Significant deviations from zero are computed with a *t*-test at the 10%, 5%, and 1% levels and are indicated by *, **, and ***, respectively.

unadjusted and market-adjusted returns for each sample grouping as well as the characteristic-based benchmark model.[5] Panel B presents adjusted returns of the market models and the three- and four-factor models where model coefficients are estimated using pre-event estimation observations. Panel C presents these same models estimated with data in the post-event estimation period.

As in BW, all the prediction models correctly predict an almost zero abnormal return when samples are randomly drawn. However, only CBBM is statistically zero, whereas the other models are significantly different than zero, except when post-event data is used to estimate the model parameters. The statistical significance does not imply economic significance however. Though the methods are statistically biased in general, the economic bias in random samples is only 0.02%. The distinction between statistical and economic bias is relevant for all of the results of this paper since finding statistical significance is not unlikely with 1000 simulations.

Across the non-random sample groupings, the ME and PR samples produce significant biases for most models using pre-event data. In particular low ME samples lead to significant positive deviations from zero and high ME samples lead to significant negative deviations from zero for all prediction models except CBBM. The multifactor models do not provide any improvement over simpler models, with all the models in Panel B finding positive abnormal returns of about 0.04% for low ME firms. Post-event data reduces the bias in Panel C to zero. Thus, there is a statistically significant size effect when samples are formed using only firms in the smallest or largest NYSE deciles using pre-event estimation data to estimate abnormal returns.

The results of Table 3 show that samples grouped by high prior returns predict 'normal' returns that are too high, leading to findings of significantly negative abnormal returns (−0.16%) for models based on pre-event data. Post-event data erases this problem but creates significant negative returns for low PR samples of about −0.04%. Low PR samples estimated with pre-event data exhibit significant positive bias. As in the small ME samples, the multifactor models do not provide substantial improvements over the market models in the PR samples. The valuation-based samples display much less bias. Only low BM firms produce significant bias among the common estimation procedures.

The biases in the PR samples are driven by a reversion to the mean between the estimation period and the event period. Prior studies of momentum have formed portfolios of past winners and losers where securities share a common calendar (Jegadeesh and Titman, 1993; Carhart, 1997). In contrast, this study groups firms into samples based on prior performance though at random dates. Thus a security in the top prior returns NYSE decile in 1973 may have a much lower average return than does a security in the same decile in 1998. The aggregation over time used in this study is appropriate to event studies, but will generate different momentum effects than will a calendar-time portfolio. Therefore, the findings in this study do not necessarily contradict or support the notion of persistence. Likewise, though the three- and four-factor models are designed to capture omitted explanatory variables, their performance in an event-study setting is not directly comparable to settings where returns share a common calendar. Moreover, the daily factor returns are from portfolios which are constructed on a yearly basis, whereas my samples of ME

[5] Unadjusted returns on day 0 do not always have the same mean as in the estimation period, since the returns are highly kurtotic and widely dispersed. Thus it is not unreasonable that the average on any one particular day should be different than the time-series average over 239 days.

Table 4
Rejection frequency of ME samples.

	Pr<0.05			Pr>0.95			Pr<0.05–Pr>0.95	
	High	Low	Difference	High	Low	Difference	High	Low
	(1)	(2)	(1)−(2)	(3)	(4)	(3)−(4)	(1)−(3)	(2)−(4)
t-statistics								
CBBM	0.049	0.052	−0.003	0.049	0.049	0.000	0.000	0.003
MEAN	0.072***	0.032***	0.040***	0.025***	0.064**	−0.039***	0.047***	−0.032***
MAEW	0.098***	0.037*	0.061***	0.025***	0.063*	−0.038***	0.073***	−0.026***
MAVW	0.052	0.031***	0.021**	0.049	0.082***	−0.033***	0.003	−0.051***
MMEW	0.087***	0.036**	0.051***	0.027***	0.068**	−0.041***	0.060***	−0.032***
MMVW	0.095***	0.034**	0.061***	0.023***	0.067**	−0.044***	0.072***	−0.033***
FF3F	0.100***	0.039	0.061***	0.020***	0.063*	−0.043***	0.080***	−0.024**
FF4F	0.101***	0.039	0.062***	0.024***	0.065**	−0.041***	0.077***	−0.026**
Standardized t-statistics								
CBBM	0.057	0.078***	−0.021*	0.054	0.055	−0.001	0.003	0.023
MAEW	0.127***	0.070***	0.057***	0.022***	0.077***	−0.055***	0.105***	−0.007
MAVW	0.061	0.044	0.017*	0.057	0.098***	−0.041***	0.004	−0.054***
MMEW	0.095***	0.043	0.052***	0.032***	0.080***	−0.048***	0.063***	−0.037***
MMVW	0.096***	0.043	0.053***	0.032***	0.080***	−0.048***	0.064***	−0.037***
FF3F	0.100***	0.047	0.053***	0.036***	0.077*	−0.041***	0.064***	−0.030**
FF4F	0.104***	0.051	0.053***	0.038***	0.084**	−0.046***	0.066***	−0.033**
Rank statistics								
CBBM	0.045	0.061	−0.016	0.061	0.042	0.019*	−0.016	0.019***
MAEW	0.080***	0.048	0.032***	0.027***	0.063*	−0.036***	0.053***	−0.015
MAVW	0.089***	0.050	0.039***	0.028***	0.065**	−0.037***	0.061***	−0.015
MMEW	0.088***	0.042	0.046***	0.021***	0.060	−0.039***	0.067***	−0.018*
MMVW	0.082***	0.048	0.034***	0.021***	0.063*	−0.042***	0.061***	−0.015
FF3F	0.084***	0.042	0.042***	0.023***	0.058	−0.035***	0.061***	−0.016
FF4F	0.088***	0.046	0.042***	0.025***	0.056	−0.031***	0.063***	−0.010
Sign statistics								
CBBM	0.045	0.059	−0.014	0.057	0.048	0.009	−0.012	0.011
MAEW	0.068**	0.044	0.024**	0.031***	0.058	−0.027***	0.037***	−0.014
MAVW	0.081***	0.046	0.035***	0.032***	0.051	−0.019**	0.049***	−0.005
MMEW	0.085***	0.040	0.045***	0.032***	0.060	−0.028***	0.053***	−0.020**
MMVW	0.076***	0.034**	0.042***	0.033**	0.064**	−0.031***	0.043***	−0.030***
FF3F	0.078***	0.038*	0.040***	0.028***	0.054	−0.026***	0.050***	−0.016*
FF4F	0.080***	0.049	0.031***	0.030***	0.054	−0.024***	0.050***	−0.005

Probability of rejection of the null hypothesis that abnormal returns equal zero, when no abnormal performance is introduced. Samples are drawn from the highest and lowest market equity deciles. Models are described in the appendix. Critical values of 0.05 and 0.95 are generated from a standard normal distribution. Numbers in parentheses under columns (1), (2), (3), and (4) indicate the *p*-value from a two-tailed exact binomial test. Numbers in parentheses under columns of differences indicate the *p*-value from a two-tailed Fisher exact test. Significance at the 10%, 5%, and 1% levels and are indicated by *, **, and ***, respectively.

and PR firms are grouped monthly. Since the CBBM model constructs a benchmark portfolio of firms from matching monthly deciles it performs better than the multi-factor models.

Though statistically significant, the degree of bias in the market and multifactor models is small in economic magnitude for most samples. However, compared to a normal daily average return of 0.1%, the −0.16% and +0.20% biases of the PR samples are quite large. In unreported tests, over a three day window, the high PR firms generate a negative bias of about −0.45% using either market or multifactor models. The longer the event window the greater will be the bias. Over a five-day window the bias is about −0.75%. Only the characteristic-based model produces no significant bias in any sample.

4.3. Statistical power of the tests

In large samples, under the null hypothesis each of the test statistics should approximate a standard normal random variable. However, the true test of a statistic is its empirical rejection frequencies. The minimization of Type I and Type II errors or, in other words, the ability to accept the null hypothesis when it is true and to reject it when it is false, are the two criteria by which a test statistic is judged. The next sections report Type I and Type II error results for each of the sample groupings, starting with samples formed by market equity.[6]

[6] Unreported results from randomly selected samples confirm findings in the event study methodology literature (BW, Corrado, 1989; Corrado and Zivney, 1992). In particular, all the prediction models generate abnormal returns insignificantly different than zero, and the non-parametric rank and sign statistics provide considerably more power than do the parametric t and standardized *t*-statistics. Performance measures of the test statistics are not reported, but are available upon request.

Table 5
Power of test statistics of ME samples.

	Abnormal performance (%)							
	Lower tail				Upper tail			
	High	ME	Low	ME	High	ME	Low	ME
Statistic/model	$-\frac{1}{2}$	-1	$-\frac{1}{2}$	-1	$+\frac{1}{2}$	$+1$	$+\frac{1}{2}$	$+1$
t-statistics								
CBBM	99.8	100.0	42.8	90.0	99.4	100.0	44.4	90.6
MEAN	99.3	100.0	42.4	90.2	98.2	100.0	52.4	94.2
MAEW	99.9	100.0	43.9	90.6	99.1	100.0	52.6	94.5
MAVW	99.9	100.0	38.5	87.8	99.8	100.0	56.4	95.6
MMEW	100.0	100.0	43.6	90.6	99.3	100.0	52.8	94.8
MMVW	100.0	100.0	43.3	90.8	99.6	100.0	52.7	94.6
FF3F	100.0	100.0	44.3	91.2	99.9	100.0	52.9	94.2
FF4F	100.0	100.0	44.1	91.2	99.7	100.0	52.5	94.3
Standardized t-statistics								
CBBM	100.0	100.0	74.0	99.8	100.0	100.0	68.4	99.7
MEAN	100.0	100.0	82.0	100.0	99.7	100.0	88.4	100.0
MAEW	100.0	100.0	83.4	100.0	100.0	100.0	83.7	100.0
MAVW	100.0	100.0	74.9	100.0	100.0	100.0	86.6	100.0
MMEW	100.0	100.0	84.5	100.0	100.0	100.0	89.8	100.0
MMVW	100.0	100.0	84.2	100.0	100.0	100.0	89.4	100.0
FF3F	100.0	100.0	84.1	100.0	100.0	100.0	89.5	100.0
FF4F	100.0	100.0	84.0	100.0	100.0	100.0	89.3	100.0
Rank statistics								
CBBM	100.0	100.0	87.7	100.0	100.0	100.0	86.8	100.0
MAEW	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
MAVW	100.0	100.0	98.8	100.0	100.0	100.0	98.2	100.0
MMEW	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MMVW	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
FF3F	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
FF4F	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Sign statistics								
CBBM	100.0	100.0	86.7	100.0	100.0	100.0	84.0	100.0
MAEW	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MAVW	100.0	100.0	99.4	100.0	100.0	100.0	99.1	100.0
MMEW	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0
MMVW	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
FF3F	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
FF4F	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0

Probability of rejection of the null hypothesis that abnormal returns equal zero for samples taken randomly from the highest and lowest ME deciles. Values in the table report the percentage of 1000 samples in which the null hypothesis of no abnormal returns is rejected. Critical values of 5% and 95% are generated from a standard normal distribution. Models are described in the appendix. Reading across each row presents the power at the 5% significance levels of each of the prediction model–statistical test combination for lower and upper tail tests.

4.3.1. Market equity samples

Table 4 presents rejection frequencies of the test statistic-prediction model combinations when no abnormal performance is artificially introduced and samples are taken from either the highest or lowest NYSE decile of market equity.[7] Unless otherwise stated, all the rejection frequencies and power measures reported are computed using a pre-event estimation period, since this is most common. Correctly specified statistics will reject the null hypothesis with a frequency equal to the nominal size of the test. Columns (1) and (3) present lower- and upper-tailed tests at the 5% level, respectively.[8]

High ME samples are incorrectly rejected by almost all of the prediction model-test statistic combinations in lower tail tests. In upper tail tests, the models accept the null of no abnormal returns too often. This is consistent with the negative mean bias in high ME samples presented in Table 3. There are no economic magnitudes here, but instead these results show that even with very little mean bias, the tests will over- or under-reject in samples of high ME firms. The over-rejection in the lower tail is substantial. The commonly employed market model-*t* statistic combination rejects the null almost twice as often as it should (0.095 vs 0.050). This finding is consistent across the test statistics, though using the value-weighted index with the *t*-statistics reduces these errors.

Skewed returns data or prediction models with biased means may lead to skewed test statistics such that the rejection frequencies are unbalanced between the tails. The last column of Table 4, reports that high ME samples yield a tendency to over-

[7] The mean-adjusted model tested with the sign and rank tests are biased by construction and produce greatly misspecified results in all sample groupings. For this reason, these results are not presented in the following tables.

[8] Tests at the 1% level also were conducted. These results are available upon request.

Table 6
Rejection frequency of PR samples.

	Pr<0.05			Pr>0.95			Pr<.05–Pr>.95	
	High	Low	Difference	High	Low	Difference	High	Low
	(1)	(2)	(1) − (2)	(3)	(4)	(3) − (4)	(1) − (3)	(2) − (4)
t-statistics								
CBBM	0.030***	0.067**	−0.037***	0.028***	0.076***	−0.048***	0.002	−0.009
MEAN	0.101***	0.013***	0.088***	0.009***	0.169***	−0.160***	0.092***	−0.156***
MAEW	0.028***	0.039	−0.011	0.027***	0.083***	−0.056***	0.001	−0.044***
MAVW	0.020***	0.028***	−0.008	0.047	0.098***	−0.051***	−0.027***	−0.070***
MMEW	0.098***	0.013***	0.085***	0.010***	0.158***	−0.148***	0.088***	−0.145***
FF3F	0.093***	0.017***	0.076***	0.009***	0.164***	−0.155***	0.084***	−0.147***
FF4F	0.096***	0.017***	0.079***	0.010***	0.164***	−0.154***	0.086***	−0.147***
Standardized t-statistics								
CBBM	0.058	0.088***	−0.030***	0.041	0.067**	−0.026***	0.017*	0.021*
MEAN	0.160***	0.014***	0.146***	0.018***	0.218***	−0.200***	0.142***	−0.204***
MAEW	0.057	0.076***	−0.019*	0.057	0.080***	−0.023**	0.000	−0.004
MAVW	0.044	0.053	−0.009	0.077***	0.091***	−0.014	−0.033***	−0.038***
MMEW	0.166***	0.012***	0.154***	0.019***	0.222***	−0.203***	0.147***	−0.210***
MMVW	0.165***	0.013***	0.152***	0.018***	0.220***	−0.202***	0.147***	−0.207***
FF3F	0.163***	0.013***	0.150***	0.018***	0.219***	−0.201***	0.145***	−0.206***
FF4F	0.168***	0.014***	0.154***	0.018***	0.211***	−0.193***	0.150***	−0.197***
Rank statistics								
CBBM	0.024***	0.109***	−0.085***	0.077***	0.029***	0.048***	−0.053***	0.080***
MAEW	0.135***	0.009***	0.126***	0.012***	0.130***	−0.118***	0.123***	−0.121***
MAVW	0.119***	0.009***	0.110***	0.016***	0.123***	−0.107***	0.103***	−0.114***
MMEW	0.133***	0.009***	0.124***	0.014***	0.107***	−0.093***	0.119***	−0.098***
MMVW	0.116***	0.010***	0.106***	0.015***	0.112***	−0.097***	0.101***	−0.102***
FF3F	0.132***	0.014***	0.118***	0.015***	0.101***	−0.086***	0.117***	−0.087***
FF4F	0.132***	0.018***	0.114***	0.018***	0.100***	−0.082***	0.114***	−0.082***
Sign statistics								
CBBM	0.028***	0.094***	−0.066***	0.079***	0.019***	0.060***	−0.051***	0.075***
MAEW	0.109***	0.012***	0.097***	0.026***	0.124***	−0.098***	0.083***	−0.112***
MAVW	0.101***	0.011***	0.090***	0.027***	0.118***	−0.091***	0.074***	−0.107***
MMEW	0.095***	0.016***	0.079***	0.019***	0.095***	−0.076***	0.076***	−0.079***
MMVW	0.084***	0.016***	0.068***	0.025***	0.097***	−0.072***	0.059***	−0.081***
FF3F	0.089***	0.019***	0.070***	0.027***	0.077***	−0.050***	0.062***	−0.058***
FF4F	0.101***	0.022***	0.079***	0.021***	0.082***	−0.061***	0.080***	−0.060***

Probability of rejection of the null hypothesis that abnormal returns equal zero, when no abnormal performance is introduced. Samples are drawn from the highest and lowest prior return deciles. Models are described in the appendix. Critical values of 0.05 and 0.95 are generated from a standard normal distribution. Numbers in parentheses under columns (1), (2), (3), and (4) indicate the *p*-value from a two-tailed exact binomial test. Numbers in parentheses under columns of differences indicate the *p*-value from a two-tailed Fisher exact test. Significance at the 10%, 5%, and 1% levels and are indicated by *, **, and ***, respectively.

reject in the lower tail compared to the upper tail. The differences columns in Table 4 show that the asymmetry between upper tail and lower tail tests as well as the asymmetric rejections between high and low ME firms is statistically significant.

Of all the models, only the CBBM model makes very few Type I errors. This model in combination with the *t*-statistic or the sign statistic is correctly specified in the size-based samples. It neither over- or under-rejects the null and is not asymmetric in the tails either. When used with a standardized *t*-statistic it slightly over-rejects in the lower tail for small firms. When used with the rank statistic it exhibits a slight asymmetry between high and low ME firms in the upper tail test, rejecting high ME firms more often than low ME firms. In summary, using standard event study methods with samples of large firms will lead to false findings of negative returns, though a characteristic-based benchmark approach does not suffer from this error.

Test statistics also must be able to reject the null when it is false. Following previous simulation studies, abnormal performance is artificially introduced into the returns data by adding a fixed return to the observed return in the amounts of −0.005, −0.010, +0.005, and +0.010. The tests are run identically as before. A test with high power should more often reject the null hypothesis for every sample simulated. The actual rejection frequencies under abnormal performance are reported in Table 5.

The more widely dispersed distributions of the low ME firms, compared to the high ME firms appears to result in lower power for low ME firms. The equal-weighted market model with a *t*-statistic rejects a positive 0.005 abnormal performance 99.3% of the time for high ME firms, but only rejects at a rate of 52.8% for low ME firms. This problem is most acute in the *t* and standardized *t* tests and is alleviated the most by using sign statistics. It is also the case that the *t* and standardized *t* tests are more likely to detect positive abnormal performance than negative abnormal performance in low ME firms. This supports the Type I results that small firms are more likely to falsely exhibit positive abnormal returns.

Power is increased tremendously by using the rank and sign tests compared to the *t* and standardized *t*. Using the MMEW model, the rank test correctly detects abnormal performance of −0.005 over twice as often in low ME firms as does the *t*-statistic (100% vs. 43.6%). Power is also improved slightly in all the test statistics with the use of multifactor models. This is evidenced by

Table 7
Power of test statistics of PR samples.

Statistic/model	Abnormal performance (%)							
	Lower tail				Upper tail			
	High	PR	Low	PR	High	PR	Low	PR
	$-\frac{1}{2}$	-1	$-\frac{1}{2}$	-1	$+\frac{1}{2}$	$+1$	$+\frac{1}{2}$	$+1$
t-statistics								
CBBM	49.0	96.0	40.5	84.5	41.5	95.1	37.1	81.4
MEAN	71.7	98.7	23.1	69.8	26.9	88.7	60.2	94.6
MAEW	51.3	96.6	39.7	83.8	46.7	96.8	41.7	86.3
MAVW	44.3	94.7	36.4	81.2	52.4	97.7	45.6	88.5
MMEW	73.2	98.9	25.3	72.5	29.4	90.8	59.4	94.7
MMVW	72.6	99.3	24.7	71.7	27.7	90.2	59.7	95.2
FF3F	73.5	99.0	26.7	72.7	29.2	90.5	60.0	94.7
FF4F	74.0	99.0	26.0	72.8	29.5	90.6	59.8	95.2
Standardized t-statistics								
CBBM	77.5	99.9	67.2	98.2	75.4	99.9	57.3	97.4
MEAN	95.8	99.7	44.9	95.7	69.5	99.9	89.4	99.9
MAEW	86.8	99.9	73.9	99.3	86.2	100.0	68.5	99.4
MAVW	81.0	100.0	68.5	99.0	89.5	100.0	73.2	99.5
MMEW	96.6	99.7	47.3	96.4	73.0	99.9	91.2	99.9
MMVW	96.8	99.7	47.6	95.9	73.4	99.9	91.6	99.9
FF3F	96.8	99.7	49.8	96.5	74.1	99.9	90.9	99.9
FF4F	97.1	99.7	49.4	96.3	75.1	99.9	91.4	99.9
Rank statistics								
CBBM	84.3	100.0	85.4	99.9	94.7	100.0	66.3	99.2
MAEW	99.9	100.0	97.1	100.0	98.1	100.0	99.8	100.0
MAVW	99.4	100.0	89.6	100.0	94.3	100.0	99.2	100.0
MMEW	100.0	100.0	99.5	100.0	99.6	100.0	100.0	100.0
MMVW	100.0	100.0	99.7	100.0	99.8	100.0	100.0	100.0
FF3F	99.9	100.0	98.5	100.0	99.3	100.0	99.7	100.0
FF4F	99.8	100.0	97.8	100.0	98.5	100.0	99.5	100.0
Sign statistics								
CBBM	80.3	100.0	82.4	99.9	91.7	100.0	60.9	98.5
MAEW	99.8	100.0	99.6	100.0	99.2	100.0	100.0	100.0
MAVW	98.9	100.0	96.5	100.0	95.6	100.0	99.6	100.0
MMEW	99.9	100.0	99.8	100.0	99.7	100.0	100.0	100.0
MMVW	100.0	100.0	99.8	100.0	99.8	100.0	100.0	100.0
FF3F	99.8	100.0	99.4	100.0	99.2	100.0	99.8	100.0
FF4F	99.9	100.0	99.3	100.0	98.1	100.0	99.4	100.0

Probability of rejection of the null hypothesis that abnormal returns equal zero for samples taken randomly from the highest and lowest prior return deciles. Values in the table report the percentage of 1000 samples in which the null hypothesis of no abnormal returns is rejected. Critical values of 5% and 95% are generated from a standard normal distribution. Models are described in the appendix. Reading across each row presents the power at the 5% significance levels of each of the prediction model–statistical test combination for lower and upper tail tests.

identical day 0 abnormal returns across models reported in Table 3, but different rejection rates. Finally, though the CBBM model is the best specified model it has the least power in general. It does however have the most symmetric power between positive and negative abnormal performance.

4.3.2. Prior returns samples

Rejection rates for samples grouped by prior returns are reported in Table 6. Almost all prediction model-test statistic combinations commit Type I errors. The models based on pre-event estimation parameters (MEAN, MMVW, MMEW, FF3F, and FF4F) over-reject in lower tail tests of samples of high PR firms and in upper tail tests of low PR firms. They also under-reject in lower tail tests for samples of low PR firms and they under-reject in upper tailed tests of high PR firms. This is consistent again with the mean bias presented above. Because returns tend to mean-revert, using the estimated parameters for firms with high prior returns leads to a prediction that is too high. Thus a normal return appears to be too low to be randomly observed and the test statistics incorrectly reject in the lower tail and do not reject often enough in the upper tail. The reverse applies to samples of firms with low prior returns.

Prediction models that do not rely on estimation periods (CBBM, MAEW, and MAVW) also exhibit errors. The market adjusted models over-reject in upper tail tests of low PR firms but under-reject in upper tail tests for high PR firms. Though CBBM also is misspecified, the deviations from the nominal size of the test are much smaller than the other models. For example, the CBBM with a *t*-statistic rejects in the upper tail for low PR firms at an empirical rate of 7.6% versus the nominal size of 5%. However, the regression-based models reject at an empirical rate of about 16% on average, more than three times the nominal size of the test.

Table 8
The effect of sample size on performance measures.

	Sample size		
	25	100	250
Panel A: Three-day abnormal return (−1; +1)			
CBBM	−0.00013	0.00017	−0.00012
MMEW	−0.00449	−0.00415	−0.00450
FF3F	−0.00455	−0.00422	−0.00453
FF4F	−0.00457	−0.00429	−0.00458
Panel B: Rejection frequency (Pr<0.05) using the rank statistic			
CBBM	0.026	0.021	0.016
MMEW	0.072	0.125	0.218
FF3F	0.086	0.124	0.233
FF4F	0.082	0.120	0.240
Panel C: Rejection frequency (Pr>0.05) using the rank statistic			
CBBM	0.062	0.086	0.105
MMEW	0.022	0.017	0.003
FF3F	0.020	0.017	0.006
FF4F	0.024	0.018	0.007

This table presents cumulative abnormal returns and the empirical rejection frequencies for a lower and upper tail test at the 5% nominal level for a three-day event window (−1; +1). The sample is from the top decile of PR. Each number reported is based on 1000 values over the sample firm averages. Models are described in the appendix.

The CBBM model also exhibits considerably less asymmetry between upper and lower tailed tests than do the other prediction models. This means that in samples of high PR firms, the CBBM rejects as frequently in upper and lower tailed tests, whereas the other models tend to reject in the lower tail ten times as often as in the upper tail. Though the economic magnitude of the mean bias produced by these methods is small, the inability to correctly identify a statistically normal return is severe.

The *t*-statistics do surprisingly well when used with the CBBM model in the PR samples compared to the nonparametric tests. However, the sign statistic leads to the least errors and produces the least asymmetry in the tails on average.

The power of the tests of PR grouped samples is presented in Table 7. Power is considerably less than in the ME samples. As in the ME samples, the CBBM measure has the lowest power but also is the most symmetric between positive and negative abnormal performance. For comparison, in high PR samples, the standard market model detects an abnormal −0.005 return in 73.2% of the simulated samples, but only detects a 0.005 abnormal return in 29.4% of the simulations. The empirical rates for CBBM are 49% and 41.5%. For most models, the values in Table 7 are asymmetric between high and low PR samples. The tests have greater power to detect negative abnormal returns for high PR firms compared to low PR firms, but less power to detect positive abnormal returns in high PR samples than low PR samples. Power is generally increased with the use of the nonparametric statistics.

4.3.3. Valuation samples

The rejection frequencies and power of the tests based on BM and EP samples do not suffer from the biases of the size- and momentum-based portfolios. Though the *t*-statistics tend to reject in the upper tail at a greater rate than in the lower tail for high BM and EP firms, in general, deviations from nominal rejection rates are small. The CBBM model again performs the best in terms of errors and symmetry between tails and high and low samples. As before, there is asymmetry in the power to detect positive versus negative abnormal performance in both BM and EP firms. The improved power of the rank and sign tests also improves symmetry between tails compared to the parametric *t* and standardized *t* tests. By a factor of about 1.5, the *t* tests are much more likely to correctly reject abnormal performance in high BM firms than in low BM firms. Thus if an event study has both high and low BM or EP firms and both experience the same abnormal performance on the event date, the *t*-tests will detect the performance in high BM firms at a much higher rate than in the low BM firms. This would lead to a finding, for instance, of a significant difference between the abnormal returns of high and low BM firms following an announcement, though both types of firms experience identical positive return increases.

4.3.4. Summary of statistical power of the tests

The results indicate that there is an economically small but statistically significant mean bias in all of the prediction models except the characteristic-based benchmark model. Furthermore, the standard procedures in event studies lead to over- and under-rejection compared to the nominal size of the test. This means that these procedures will lead to false findings of significantly negative returns in samples of large firms with high prior returns, such as firms making acquisitions or initiating dividends. Alternatively, small firms with low prior returns will appear to have significantly positive returns when none actually exist.

4.4. Event period variance increases

As has been documented in previous literature, an event period variance increase may cause incorrect rejection rates when no abnormal returns are present. To analyze how each prediction-model test statistic combination is affected by a day 0 variance

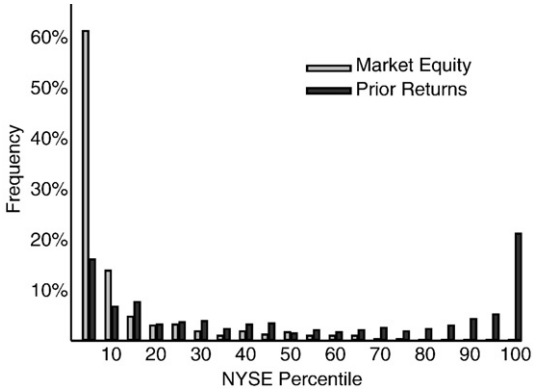

Fig. 1. Size and return distributions of reverse-split stocks. This figure presents the relative frequency of stocks that reverse-split between 1972 and 2002 in each 5th percentile of the NYSE for market equity and 1-year prior returns. Only stocks with split ratios of −0.2 or less and with a stock price of at least $1 the day before the *ex* date are included.

increase, event day return variances are artificially increased following the procedure in BW, where day 0 variance is increased using a day 5 return, but adjusted such that the mean is the same.[9]

In unreported results, variance increases lead to significant over-rejection of the null hypothesis in both upper and lower tailed tests by the *t* and standardized *t*-statistics for all prediction models and all sample groups, including random samples. These rejection rates are quite high, around seven times the correct nominal rate of 5%. The sign test performs better than the rank test and both perform much better than the *t*-tests, though they still commit serious errors.

4.5. Sample size and model performance

Table 8 presents the bias in means and the rejection frequencies for samples of various sizes. The estimates are for a three day window around the announcement date $(-1, +1)$ where samples are drawn from the top decile of prior returns. If the mean bias or the Type I and II errors reported above are caused by small samples, we should see improvement as the sample size increases. Instead, omitted variable bias will not diminish if the sample size is increased.

The mean bias is virtually unchanged whether the sample size is 25, 100, or 250 stocks for all models. Only CBBM seems to have variability in the estimate of the mean, but no clear relationship with sample size is obvious. In contrast, the rejection frequencies change substantially as the sample size increases. For the models using pre-event estimation periods, the over-rejection in lower tail tests increases substantially, as does the under-rejection in the upper tail tests. This is because the statistics are becoming more precise, though they are centered about a biased mean. The CBBM model performs the best because it has the least mean bias. For comparison, when the sample size is 25, the CBBM model rejects 2.6% of the time in the 5% lower tail tests compared to the FF4F model which rejects 8.2% of the time. When the sample size is 250, the CBBM model rejects slightly less often (1.6%) versus the large increase in the FF4F rejection rate of 24%. Again, the economic magnitude is small but the measures of statistical significance change dramatically.

5. Applications to actual event samples

The results presented to this point are generated by extreme marginal distributions of the four pricing characteristics (ME, PR, BM, and EP). In actual events, samples will have joint distributions across these characteristics, and abnormal returns will be aggregated over a longer event window.

5.1. Reverse stock splits

Motivations for reverse stock splits are to comply with minimum price requirements mandated by exchanges (Martell and Webb, 2008), to appeal to institutional investors, and to reduce transaction costs (Han, 1995). Since stocks that reverse split are likely to have had low returns in the prior year and are typically small, they provide a good setting to test the potential biases found above in an actual event. Using the distributions of size and returns from actual stocks that reverse split, I simulate event studies as before and investigate the magnitude of biases.

Following Martell and Webb (2008), data are taken from the CRSP daily file from 1972 to 2002. To be included in the sample the reverse split ratio must be less than −0.2, or 4:5, and the price of the stock the day before the *ex*-date must be $1 or more. In

[9] This method of artificially increasing the variance will produce other unintended changes in the shape of the returns distribution which will affect the rejection rates. Also, this method of increasing the variance will produce average return changes if the time series average return is not an accurate prediction of the day 0 event return. This was the case most noticeably for the prior return samples analyzed previously. Thus incorrect rejection in prior return samples using the transformation above may be due to average return changes and not variance increases.

addition there must be at least 50 non-missing observations in the estimation period (−239, −6), and all observations must be present in (−15, +5). This generates 564 reverse stock splits. The distributions of market equity and prior year returns are presented in Fig. 1. More than 60% of these firms are smaller or equal to the 5th percentile of NYSE firms. However, the prior returns are quite low for many firms, but others actually have high returns compared to the NYSE. Because NYSE firms are larger and have lower returns on average than NASDAQ firms, these two results are not surprising.

Next, I simulate an event study where firms are chosen randomly to match the distribution of size and prior returns presented in Fig. 1. For each firm in the actual split sample, I randomly select a matching firm on the same date from the same bivariate distribution of size and prior returns where the distributions are over deciles of the NYSE level. I repeat this simulated event study of 564 firms 500 times.

The one-day abnormal return from the FF3F model in the simulation is 0.007% and the 3-day abnormal return is 0.02%. The other models' CARs are similar in magnitude. Of the four test statistics, only the rank statistic identifies that this bias is significant. With no abnormal performance introduced, the test statistics reject at a rate roughly equal to the nominal size of the test in the lower tail, but the *t* and standardized *t* statistics over-reject in the upper tail, with the worst performance by the standardized *t*-statistic. The rank and sign tests do not suffer from this misspecification.

The event study returns surrounding the actual split *ex*-dates are −2.22% for a 1-day window using the FF3F model. For a five-day window (−2, +2) the CAR is −2.46%. The other models generate very similar results. Thus the biases in this case are not large compared to the actual CARs. If the returns were small or positive for a sample distributed similarly to the reverse split sample, the biases would have a greater impact. The following demonstrates this by forming hypothetical distributions constructed to resemble samples of firms that have distributions across the pricing characteristics similar to firms in a variety of actual events.

5.2. Other simulated samples

A simulation of 1000 samples of 250 firms taken from the top 25th NYSE percentile of market equity and prior returns and the bottom 25th percentile of book-to-market is performed where no abnormal performance is introduced. This distribution is designed to resemble a sample of large acquirers or firms forward-splitting their stock. CARs generated from simulations over a three-day period surrounding the event date (−1, +1) produce an abnormal return of −0.327% using the equally weighted market model. The actual three-day CAR reported in Moeller et al. (2004) for large acquirers is 0.076%. Correcting for the sample selection bias, the $CAR_{(-1,\ +1)}$ would be 0.403% for large acquirers, a six-fold increase in the prior result. Though the bias in abnormal returns is small, the bias in dollar returns is quite large. At the 85th percentile of NYSE market equity ($13.8 billion in 2007), the three-day bias of −0.327% generates a negative abnormal dollar return of approximately $45 million. Thus small return biases may lead to large dollar biases in event studies of large firms. In addition, rejection rates in the lower tail for a 5% test are 16.3% using a *t*-statistic, over three times the nominal size of the test. However, as in the reverse split case, if the announcement returns are large, the forecast bias will be relatively insubstantial. The five-day CAR using the MAVW model for stock splits reported in Rankine and Stice (1997) is 1.44%, compared to a simulated bias of only 0.13%.

A second simulation was performed to resemble small acquirers, firms announcing new exchange listings, and firms making seasoned equity offerings. These firms were taken from the bottom 25th percentile of NYSE market equity and book-to-market, and the upper 25th percentile of prior returns. Abnormal returns with no abnormal performance for this sample are also biased downward, with a three-day CAR of −0.55% reported using the equally weighted market model. The negative high prior return bias outweighs the positive small firm bias and produces negative deviations from zero. The rejection frequencies are significantly different from their nominal sizes, rejecting in over 20% of the simulated samples in lower tail 5% tests and in only 1% of the simulated samples for upper tail tests. The characteristic-based benchmark model again produces the least mean bias and better rejection frequencies.

A third simulation was performed where samples were drawn from the bottom 5th NYSE percentile of size and prior returns. This distribution is chosen to resemble the potential sample of an event study investigating distressed firms. In this case, simulated three-day CARs with no abnormal performance yield positive and significant abnormal returns of 0.93% for the market and multifactor models, suggesting that abnormal returns reported using standard methods may be too high for a sample of distressed firms. In contrast, the benchmark model exhibits no significant bias. Rejection rates are significantly different than their nominal levels in lower tail tests (less than 1% empirical rejection rate) and excessively high in upper tail tests (30% approximately). Thus significant and positive abnormal returns may be found for this sample where none actually exist at levels that are economically significant. These three simulations show that mean bias and over- and under-rejection are found in standard methods even for samples that are not taken from extreme deciles, but rather have firm characteristics similar to actual event samples.

6. Conclusion

This paper conducts simulations of event studies where sample securities are grouped by the common characteristics of market equity, prior returns, book-to-market, and earnings-to-price ratios using daily returns from 1965 to 2003. A battery of prediction models and test statistics are compared for possible null rejection biases when returns are expected to have zero abnormal performance, when returns are artificially increased and decreased, and when variance is artificially increased.

In support of BW, when samples are randomly drawn, all the prediction models generate abnormal returns with only minor differences from zero with correct rejection rates in general. In contrast to the findings of BW, many of the prediction models are statistically misspecified for non-random samples grouped by prior returns and market equity. Specifically, the commonly used

OLS market model with a *t*-test produces incorrect rejection rates under no abnormal performance for securities that are grouped by size, prior returns, and book-to-market ratios. These rejection rate errors are driven by false statistically significant positive returns for samples characterized by small firms and firms with low prior returns. Moreover, the power of the *t*-test to detect abnormal performance is lower than the nonparametric tests and displays considerable bias. However, since the nonparametric tests compare medians rather than means, their appropriate use will be dictated by the context of an event study.

Furthermore, the use of multifactor models does not decrease the forecast error bias compared to simpler methods. Instead only the characteristic-based benchmark approach exhibits no mean bias in any sample grouping. It does however over- and under-reject in samples grouped by prior returns, though not to the degree that the other models do. Table 3 also shows that using post-event data to estimate model coefficients can significantly reduce the positive small firm bias for samples grouped by size and the negative bias found in samples characterized by high prior returns. Thus, when conducting an event study it is recommended to choose the estimation period that is the most 'normal' and that produces the least bias. Though specification error is reduced using post-event data for the random event dates in this paper, post-event estimation periods will only make sense in some event studies.

The findings presented here suggest that it is incorrect to generalize the random sample results of BW to the non-random samples characteristic of actual event studies. Thus, announcement day abnormal returns found in prior research may be biased by statistical error in estimating normal returns. The economic consequences of the bias will depend upon the relative size of the abnormal returns found in an event study. The bias will become more economically significant in certain cases where the compound bias is increasing, such as a sample of distressed firms characterized by simultaneously low ME, low PR, and high BM values.

Finally, though announcement day abnormal return biases may be small, when samples are comprised of large firms, dollar abnormal return biases may be large. Even for small firms the statistical biases will be important when researchers look for cross-sectional explanations of abnormal returns, which are typically small but significant differences. For example, Lipson et al. (1998) report higher abnormal returns for NASDAQ firms initiating dividends compared to NYSE firms (1.44% vs. 0.76%). This difference may be explained in part by estimation bias, rather than unexplained investor sentiment and may in fact be statistically insignificant.

Appendix A

A.1. Prediction models

1. Mean Adjusted Return (MEAN)

$$A_{i,t} = R_{i,t} - \overline{R_i} \tag{A.1}$$

where $$\overline{R_i} = \frac{1}{239}\sum_{t=-244}^{-6} R_{i,t} \tag{A.2}$$

2. Market Adjusted Return: Equal-Weighted (MAEW) and Value-Weighted (MAVW)

$$A_{i,t} = R_{i,t} - R_{M,t}, \tag{A.3}$$

where $R_{M,t}$ is the return on the CRSP equal-weighted or value-weighted index for day *t*.

3. Market Model: Equal-Weighted (MMEW) and Value-Weighted (MMVW)

$$A_{i,t} = R_{i,t} - \hat{\alpha}_i - \hat{\beta}_i R_{M,t}, \tag{A.4}$$

where R_M is the CRSP equal-weighted or value-weighted index and $\hat{\alpha}$ and $\hat{\beta}$ are OLS parameter estimates from the estimation period.

4. Fama–French 3 Factor Model (FF3F)

$$A_{i,t} = R_{i,t} - \hat{\alpha}_i - \hat{\beta}_i R_{M,t} - \hat{s}_i \mathrm{SMB}_t - \hat{h}_i \mathrm{HML}_t, \tag{A.5}$$

where R_M is the CRSP value-weighted index, SMB (Small Minus Big) is a mimicking portfolio to capture risk related to size, and HML (High Minus Low) is a mimicking portfolio to capture risk associated with book-to-market characteristics. The coefficient estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{s}$, and $\hat{h}$ are obtained from an OLS regression on the estimation period returns. See Fama and French (1993) for more details on the risk factors.

5. Carhart 4 Factor Model (FF4F)

$$A_{i,t} = R_{i,t} - \hat{\alpha}_i - \hat{\beta}_i R_{M,t} - \hat{s}_i \mathrm{SMB}_t - \hat{h}_i \mathrm{HML}_t - \hat{u}_i \mathrm{UMD}_t, \tag{A.6}$$

where R_M, SMB, and HML are the same as above and UMD (Up Minus Down) is a mimicking portfolio designed to address risk associated with prior returns by subtracting a portfolio of low prior return firms from a portfolio of high prior return firms where prior returns are measured over the months $t-12$ to $t-2$. The daily risk factors, SMB, HML, and UMD were generously provided on Kenneth French's Web site.

6. Characteristic-Based Benchmark Model (CBBM)

$$A_{i,t} = R_{i,t} - R_{\mathrm{BM},t}, \tag{A.7}$$

where $R_{\mathrm{BM},t}$ is the return on the equal-weighted portfolio of a sample of stocks that share the same size and one year prior return NYSE deciles as stock i at time t. For this paper I use a benchmark sample of 10 stocks.

A.2. Test statistics

1. t-statistic

The day 0 t-statistic is computed as described in BW,

$$t \text{ statistic} = \frac{\overline{A_0}}{\hat{S}\left(\overline{A_t}\right)}, \tag{A.8}$$

where

$$\overline{A_t} = \frac{1}{N_t}\sum_{i=1}^{N_t} A_{i,t} \tag{A.9}$$

$$\hat{S}\left(\overline{A_t}\right) = \sqrt{\frac{1}{238}\sum_{t=-244}^{t=-6}\left(\overline{A_t} - \overline{\overline{A}}\right)^2} \tag{A.10}$$

$$\overline{\overline{A}} = \frac{1}{239}\sum_{t=-244}^{t=-6}\overline{A_t}, \tag{A.11}$$

and N_t is the number of sample securities whose excess returns are available at date t.

2. Standardized t-statistic

The standardized t-statistic accounts for heteroskedasticity in abnormal returns across firm returns by a standard deviation normalization. The day 0 statistic is computed as follows,

$$\text{standardized } t\text{-statistic} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{A_{i,0}}{\hat{S}(A_i)} \tag{A.12}$$

where

$$\hat{S}(A_i) = \sqrt{\frac{1}{T_i - 1}\sum_{t=-244}^{-6}\left(A_{i,t} - \overline{A_i}\right)^2} \tag{A.13}$$

T_i = Number of non-missing returns for firm i in the estimation period

$$\overline{A_i} = \frac{1}{T_i}\sum_{t=-244}^{-6} A_{i,t} \tag{A.14}$$

3. Rank statistic

The rank test is computed as in Corrado and Zivney (1992),

$$\text{Rank statistic} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\left(U_{i0} - \frac{1}{2}\right)}{S(U)} \tag{A.15}$$

where

$$U_{it} = \frac{\mathrm{rank}(A_{it})}{(1 + M_i)}, \quad t = -244, \ldots, +5 \tag{A.16}$$

A_{it} excess return of security i on day t
M_i number of non-missing returns for security i
N number of securities in the sample portfolio

$$S(U) = \sqrt{\frac{1}{250}\sum_{t=-244}^{t=+5}\left(\frac{1}{\sqrt{N_t}}\sum_{i=1}^{N_t}\left(U_{it} - \frac{1}{2}\right)\right)^2} \tag{A.17}$$

N_t = number of non-missing returns in the cross-section of N-firms on day t

4. Sign statistic

The sign test is also computed as in Corrado and Zivney (1992),

$$\text{Sign statistic} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{G_{i0}}{S(G)} \tag{A.18}$$

where

$$G_{it} = \text{sign}(A_{it} - \text{median}(A_{it})) \quad t = -244, \ldots, +5 \tag{A.19}$$

$\text{sign}(x) = -1, +1$, or 0

$$S(G) = \sqrt{\frac{1}{250}\sum_{t=-244}^{t=+5}\left(\frac{1}{\sqrt{N_t}}\sum_{i=1}^{N_t}G_{it}\right)^2} \tag{A.20}$$

References

Agrawal, A., Jaffe, J.F., Mandelker, G.N., 1992. The post-merger performance of acquiring firms: a re-examination of an anomaly. J. Finance 47, 1605–1621.
Alderson, M.J., Betker, B.L., 2006. The specification and power of tests to detect abnormal changes in corporate investment. J. Corp. Finance 12, 738–760.
Asquith, P., 1983. Merger bids, uncertainty, and stockholder returns. J. Financ. Econ. 11, 51–83.
Banz, R.W., 1981. The relationship between return and market value of common stocks. J. Financ. Econ. 9, 3–18.
Barber, B.M., Lyon, J.D., 1996. Detecting abnormal operating performance: the empirical power and specification of test statistics. J. Financ. Econ. 41, 359–399.
Basu, S., 1983. The relationship between earnings' yield, market value and return for NYSE common stocks: further evidence. J. Financ. Econ. 12, 129–156.
Brav, A., 2000. Inference in long-horizon event studies: a bayesian approach with applications to initial public offerings. J. Financ. 55, 1979–2016.
Brav, A., Geczy, C., Gompers, P.A., 2000. Is the abnormal return following equity issuances anomalous? J. Financ. Econ. 56, 209–249.
Brown, S.J., Warner, J.B., 1985. Using daily stock returns: the case of event studies. J. Financ. Econ. 14, 3–31.
Brown, S.J., Goetzmann, W.N., Ross, S.A., 1995. Survival. J. Finance 50, 853–873.
Campbell, J.Y., Lo, A.W., MacKinlay, A.C., 1997. The Econometrics of Financial Markets. Princeton University Press.
Campbell, J.Y., Hilscher, J.D., and Szilagyi, J., 2005, In search of distress risk, Harvard Institute of Economic Research Discussion Paper No. 2081.
Carhart, M.M., 1997. On persistence in mutual fund performance. J. Finance 52, 57–82.
Copeland, T.E., Mayers, D., 1982. The Value Line enigma (1965 – 1978): a case study of performance evaluation issues. J. Financ. Econ. 10, 289–321.
Corrado, C.J., 1989. A nonparametric test for abnormal security-price performance in event studies. J. Financ. Econ. 23, 385–395.
Corrado, C.J., Zivney, T.L., 1992. The specification and power of the sign test in event study hypothesis tests using daily stock returns. J. Financ. Quant. Anal. 27, 465–478.
Daniel, K., Grinblatt, M., Titman, S., Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. J. Finance 52, 1035–1058.
Dharan, B.G., Ikenberry, D.L., 1995. The long-run negative drift of post-listing stock returns. J. Finance 50, 1547–1574.
Dimson, E., Marsh, P., 1986. Event study methodologies and the size effect: the case of UK press recomendations. J. Financ. Econ. 17, 113–142.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 3–56.
Fama, E.F., French, K.R, 1996. Multifactor explanations of asset pricing anomalies. J. Finance 51, 55–84.
Fama, E.F., French, K.R, 1997. Industry costs of equity. J. Financ. Econ. 43, 153–193.
Fama, E.F., 1998. Market efficiency, long-term returns, and behavioral finance. J. Financ. Econ. 49, 283–306.
Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. J. Finance 47, 427–465.
Gregory, A., 1997. An examination of the long run performance of UK acquiring firms. J. Bus. Finance Account. 24, 971–1002.
Gur-Gershgoren, G., Hughson, E., Zender, J., 2008. A simple-but-powerful test for long-run event studies. Robert Day School of Economics and Finance Research paper No. 2008–8.
Han, K.C., 1995. The effects of reverse splits on the liquidity of the stock. J. Financ. Quant. Anal. 30, 159–169.
Ikenberry, D.L., Rankine, G., Stice, E.K., 1996. What do stock splits really signal? J. Financ. Quant. Anal. 31, 357–375.
Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. J. Finance 48, 65–91.
Kothari, S., Warner, J.B., 2005. Handbook of corporate finance: empirical corporate finance chapter. Handbooks in Finance Series, Elsevier North Holland.
Lipson, M.L., Maquieira, C.P., Megginson, W., 1998. Dividend initiations and earnings surprises. Financ. Manage. 27, 36–45.
Loughran, T., Ritter, J.R., 2000. Uniformly least powerful tests of market efficiency. J. Financ. Econ. 55, 361–389.
Lund, R.E., Lund, J.R., 1983. Algorithm AS 190: probabilities and upper quantiles for the studentized range. Applied Statistics 32, 204–210.
Lyon, J.D., Barber, B.M., Tsai, C., 1999. Improved methods for tests of long-run abnormal stock retirns. J. Financ. 54, 165–201.
Mandelker, G., 1974. Risk and return: the case of merging firms. J. Financ. Econ. 1, 303–335.
Martell, T.F., Webb, G.P., 2008. The performance of stocks that are reverse split. Rev. Quant. Financ. Account. 30, 253–279.
Michaely, R., Thaler, R.H., Womack, K.L., 1995. Price reactions to dividend initiations and omissions: overreaction or drift? J. Finance 50, 573–608.
Mitchell, M.L., Stafford, E., 2000. Managerial decisions and long-term stock price performance. J. Bus. 73, 287–329.
Moeller, S.B., Schlingemann, F.P., Stulz, R.M., 2004. Firm size and the gains from acquisitions. J. Financ. Econ. 73, 201–228.
Pearson, E., Hartley, H. (Eds.), 1966. third edn. Biometrika Tables for Statisticians, vol. I. Cambridge University Press, Cambridge.
Rankine, G., Stice, E.K., 1997. The market reaction to the choice of accounting method for stock splits and large stock dividends. J. Financ. Quant. Anal. 32, 161–182.
Rhodes-Kropf, M., Robinson, D.T., Viswanathan, S., 2005. Valuation waves and merger activity: the empirical evidence. J. Financ. Econ. 77, 561–603.
Schwert, G.W., 2000. Hostility in takeovers: in the eyes of the beholder? J. Finance 55, 2599–2640.
Wooldridge, J.M., 2000. Introductory Econometrics: A Modern Approach. South-Western College Publishing.