

A Simple-But-Powerful Test for Long-Run Event Studies^{*}

Gitit Gur-Gershgoren
Israel Securities Authority

Eric Hughson
Claremont McKenna College

Jaime F. Zender
University of Colorado at Boulder

First Draft: September, 2005
Current Draft: October 12, 2008

^{*} Preliminary and incomplete, please do not quote without authors' permission. We thank Steve Manaster and seminar participants at the University of Illinois, University of North Carolina, University of Oklahoma, University of California at San Diego, and Vanderbilt University for helpful discussions.

A Simple-But-Powerful Test for Long-Run Event Studies

Abstract

Testing for long-run abnormal performance has become an important part of the finance literature. We propose a test for abnormal performance in long-run event studies using the buy and hold abnormal return (*BHAR*). We augment the control firm approach of Barber and Lyon (1997) by using multiple control firms per event firm to create multiple correlated *BHARs* for each event firm. Using the control firm structure allows us to easily avoid the new listing, rebalancing, and skewness biases. Further, despite the induced correlation amongst the *BHARs*, using multiple control firms allows us to increase the power of the test beyond that of existing tests. Most importantly, we show that, with the appropriate choice of control firms, our test is well-specified in both random and nonrandom samples.

Introduction

Long-run event studies, which have been used to examine the price behavior of equity for periods of one to five years following significant corporate events (e.g. IPOs, SEOs, repurchases, bond rating changes, etc.) are an increasingly important part of the finance literature. Despite considerable interest in the long-run behavior of prices relative to expectations, finance scholars are engaged in a continuing debate concerning the appropriate measure of long-run abnormal performance and the appropriate statistical methodology to use to test for the significance of any measured abnormal performance.

Here we develop a well-specified and powerful test for long-run abnormal performance as measured by the buy and hold abnormal return (*BHAR*); defined as the difference between the long-run return for a sample asset and that of a benchmark asset selected to capture expected return. The buy and hold abnormal return is the focus of this study as it provides a measure of long-run investor experience, the focus of most long-run event studies (see for example Ritter (1991) or Loughran and Ritter (1995)). Our approach modifies the control firm approach (Barber and Lyon (1997)) by using multiple performance measures per event firm to address the associated power problem. The resulting test is uniformly more powerful than those currently extant and, more importantly, is well specified in all samples, random and non-random, we have examined.

The most popular alternative abnormal return measures, the cumulative abnormal return (*CAR*) or average abnormal return (*AAR*), measure average periodic abnormal returns and so are biased estimators of long-run investor experience. Fama (1998) and Mitchell and Stafford (2000) have argued for their use and for calendar time tests for the presence of abnormal performance primarily because of statistical problems associated

with the use of the *BHAR* and the associated test statistics. Our approach is designed to address these concerns and so allows for the legitimate use of the *BHAR*.

Barber and Lyon (1997) and Lyon, Barber, and Tsai (1999) identify three problems with inference in long-run event studies using the *BHAR*. Labeling these problems the new listing, rebalancing,¹ and skewness biases, they use simulations to examine the impact of these biases on inference when abnormal performance is measured using the *BHAR* and standard tests (predominantly t-statistics²) are applied. When a reference portfolio is used to capture normal or expected return, the new listing and rebalancing biases can be addressed in a relatively simple way by careful construction of the reference portfolio (see Lyon, Barber, and Tsai (1999)).

The more serious problem associated with the use of a reference portfolio to capture expected return is the skewness bias. This bias arises because the long-run return of a portfolio is compared with the long-run return of an individual asset. The long-run return of an individual security is highly skewed; whereas the long-run return for a reference portfolio (due to diversification) is not. Consequently, the *BHAR*, the difference between these returns, is also skewed. Barber and Lyon demonstrate in simulations that the *BHAR*'s positive skewness causes standard tests to have the wrong size (the null hypothesis is rejected too often when it is true, see also Kothari and Warner (1997)) and causes the power of the test to be asymmetric; rejection rates are far higher when induced abnormal returns are negative than when they are positive.

The skewness bias does not arise when a control firm rather than a reference portfolio is used as the long-run return benchmark. In that case, the *BHAR* is measured as

¹ See Canina, Michaely, Thaler, and Womack (1996).

² An alternative, Bayesian approach is presented in Brav (2000).

the difference between the long-run holding-period returns of the event firm's equity and that of a control firm. Although the distribution of each asset's holding-period return is highly skewed, the distribution of their difference is not.³ Therefore, standard statistical tests based on the control firm approach have the right size in random samples.

Unfortunately, Barber and Lyon (1997) show that the power of standard tests based on the control firm approach is very low when compared with the reference portfolio approach. Simply put, the use of a control firm is a noisier way to control for expected returns than is the use of a reference portfolio and this added noise reduces the power of the test. The variance of the difference between the returns on two individual assets is generally much higher than the variance of the difference between an asset's return and that of a portfolio; even when the control firm is chosen carefully. Hence, for the control firm approach, powerful tests require very large samples.

In order to develop a well-specified test with high power, Lyon, Barber, and Tsai (1999) examine two ways to modify the reference portfolio approach to fix the associated size problem: the use of p-values generated from the empirical distribution of long-run abnormal returns and the use of skewness-adjusted t-statistics.⁴ These methods, combined with careful construction of reference portfolios to remove the rebalancing and new listing biases, solve the size problem in "random" samples. These corrections, however, do not yield well-specified tests in many of the non-random samples considered

³ Barber and Lyon (1997, Table 7) report skewness for annual *BHARs* using the control firm approach of about 0.4 compared to 8.0 when the reference portfolio approach is used.

⁴ In a strict statistical sense the comparisons made in Lyon, Barber and Tsai (1999) (and in the present work) are not standard comparisons of the power to reject the null across different tests. The different methods of computing *BHARs* define normal returns differently. Changing the definition of normal returns changes slightly the null hypothesis, because the definition of abnormal returns is also altered. However, the different ways to construct the test, i.e. the use of pseudoportfolios to calculate empirical p-values, the use of skewness-adjusted t-statistics or our use of multiple *BHARs* per event firm, simply create different statistical interpretations of the same underlying economic hypotheses: *is there long-run abnormal performance after a specific event?* It is in this context that the comparisons are considered.

by Lyon, Barber, and Tsai. In non-random samples the use of a standard reference portfolio approach often fails to match the expected return of the event firm with the expected return of the reference portfolio resulting in a mis-specified test (the “bad model” problem highlighted by Fama (1998)). Furthermore, when the return on a diversified portfolio is used to capture expected returns, there is no offset of any contemporaneous correlation of idiosyncratic returns that may exist across firms. This problem is likely to be heightened when events are highly clustered in time.

Rather than refine the reference portfolio approach, we introduce a methodology that increases the power of the control firm approach by matching each of N event firms with M control firms. This results in M correlated *BHARs* per event firm and we use a Wald test statistic to account for this induced correlation when testing whether the average of the NM *BHARs* is zero. We can, in this way, retain the benefits of the control firm approach and at the same time greatly increase the power of the test. By careful choice of control firms, our methodology provides a test whose power is uniformly higher than that of any examined by Lyon, Barber, and Tsai (1999). More importantly, the flexibility provided by the selection of a relatively few control firms per event firm provides well specified tests in all the samples, random and non-random, we consider.

Although a variety of possibilities exist, we examine the properties of two matching procedures. The first matches each event firm with multiple control firms based on a set of characteristics that have been shown to be related to average returns (size, book-to-market, beta, etc.). The second is a statistical (“maximal R^2 ”) matching procedure. From a set of K potential control firms, we choose the M that jointly maximize the R^2 of the regression of an event firm’s return on its set of M control firm

returns over a pre-event estimation period. The K potential control firms are selected to minimize the possibility that the maximal R^2 procedure will result in problems of “overfitting.” While both approaches provide tests with the correct size, we find that the maximal R^2 procedure consistently produces a more powerful test. Using either approach, the power of the tests increase significantly until $M = 3$ control firms per event firms are used. Further increases in M result in relatively small increases in power.

We then investigate the performance of “conditional” versions of our matching procedures on non-random samples. Lyon, Barber, and Tsai (1999) show that the reference portfolio approach often leads to mis-specified tests when the event firms share certain firm characteristics. The underlying problems are either that the expected return of the event firms is not well-matched by the expected return of standard reference portfolios when the “event” affects a non-random sample of firms or that idiosyncratic returns are correlated across event firms. Because our approach uses control firms to capture normal returns and relatively few control firms per event firm, we modify our matching procedures to capture expected returns sufficiently well for non-random samples of event firms that they provide well-specified tests.

Finally, we investigate the effect of calendar clustering (see Fama (1998) or Mitchell and Stafford (2000)) of events. The potential problem here is that the observed positive contemporaneous cross-sectional correlation in idiosyncratic returns suggests that samples clustered in calendar time may generate positive cross-sectional correlation in the *BHARs*, which in turn would cause over-rejection of the null hypothesis when it is true. The use of control firms rather than a reference portfolio for defining abnormal performance implies that the idiosyncratic component of the measure of abnormal

performance is the difference in the contemporaneous idiosyncratic returns of two similar assets and so some of the observed correlation in idiosyncratic returns may be removed by the model used to adjust for expected returns (Fama (1998)). Empirically, we show that our procedures generate *BHARs* that are only minimally correlated across event firms and result in test statistics that have the right size even when events are maximally clustered in simulations so that they all begin on the same day.

Our paper is organized as follows. Section 1 describes our methodology. In section 2, we describe the data generation procedure for our simulations. In section 3, we provide our results and a comparison with those in earlier studies. Section 4 concludes.

1. Methodology

1.1 The Abnormal Return Metric

A fundamental choice for any long-run event study concerns the measure of long-run abnormal return. The literature concentrates on the use of either compounded long-run abnormal returns (BHARs) or measures of average periodic performance (CARs or AARs). While BHARs directly measure investor experience, the measure of concern for most long-run event studies, Fama (1998) raises several significant concerns with their use as a performance metric. First, the “bad model” problem (errors in specifying expected return) is more of a problem for the use of BHARs than it is for the use of average periodic abnormal return measures due to compounding. Secondly, long-run BHARs (under the reference portfolio approach) are highly skewed, causing standard tests to have the wrong size. Finally, to the extent that events tend to be clustered in time, cross-sectional correlation of contemporaneous returns may also cause tests using

BHARs to be mis-specified. In addition to the bad model problem, if the idiosyncratic components of returns are correlated in the cross-section⁵ BHARs may also be cross-sectionally correlated. In that case, estimates of BHAR variation that do not account for the correlation between contemporaneous observations will be understated (see Brav (1997)) causing the associated test statistics to reject the null too often. Individually, each of these problems would cause one to distrust hypothesis testing done on BHARs. Jointly they motivate Fama to argue for the use of CARs.

The main argument against using average periodic abnormal returns in long-run event studies is that they do not measure the variable of interest. Barber and Lyon (1997) show that CARs are a biased predictor of investor experience. Therefore a failure to reject the null of no abnormal return measured as average period abnormal return does not imply a lack of abnormal return as measured by the BHAR. It is therefore useful to develop well-specified tests of abnormal return using buy and hold abnormal returns as the performance metric. Any such test must, of course, address Fama's (1998) concerns.

Our approach alleviates each of Fama's concerns. As discussed in the introduction, we modify the control firm approach by matching each event firm with several control firms. Therefore, by design (see Barber and Lyon (1997) table 8) skewness in the BHARs is much less of a problem as compared to the reference portfolio approach. Barber and Lyon also show that the control firm approach leads to well-specified tests using conventional t-statistics.

As noted by Fama (1998), errors in specifying expected returns are compounded when using BHARs as the measure of abnormal performance which could lead to over-

⁵ See Mitchell and Stafford (2000), Collins and Dent (1984), Sefcik and Thompson (1986) and Bernard (1987).

rejection of the null if the errors are significant. Furthermore, as Fama notes, the bad model problem is likely to be most problematic when using equally weighted abnormal returns. This is because small stocks have been shown to provide the greatest difficulty in capturing expected returns and are weighted relatively heavily when returns are equally-weighted. Although our methodology is easily adapted to any weighting scheme, to determine whether equal weighting presents a problem, we consider only equally-weighted BHARs in our simulations. Despite this disadvantage, our simulations show that our approach is well specified in both random and non-random samples of event firms. Empirically, therefore, we find that our approach, with its associated procedures for selecting control firms, leads to firm specific proxies for normal or expected returns that are accurate enough that the bad model problem is not a significant concern.

Finally, if contemporaneous idiosyncratic returns are cross-sectionally correlated, then even if the systematic component of returns is captured accurately, the cross-sectional correlation of the residuals could also lead to a mis-specified test if it is not accounted for in the measure of BHAR variation. Mitchell and Stafford (2000) report average cross-sectional correlation in monthly BHARs (using the reference portfolio approach) between 0.002 and 0.0177 when there is complete calendar time overlap. They also show that ignoring correlations even of this magnitude can lead to a sufficient underestimation of BHAR variance so standard tests significantly over-reject the null.

The reference portfolio approach uses a large, well-diversified portfolio's return to proxy for the systematic component of returns. Therefore, any cross-sectional correlation in residual returns will be inherited by the BHARs. The situation is different with the control firm approach. In this case the BHAR is the difference between the

returns on two individual assets and so (if the systematic components of returns for the assets are the same) the BHAR is the difference between their residual returns. Therefore cross-sectional correlation in residual returns does not necessarily imply cross-sectional correlation in the BHARs. Indeed, in our simulations average cross sectional correlation between monthly BHARs, analogous to that reported by Mitchell and Stafford, is .00015. This measure is generated assuming *all* events occur on the same day and is an order of magnitude smaller than that reported by Mitchell and Stafford (2000). Furthermore, when we cluster events in time so that they all occur on the same day (to induce the maximal amount of cross-sectional correlation) our approach yields a well-specified test.

1.2 The Model

Consider an “event” shared by a sample of N firms, $i=1, \dots, N$. We use the *BHAR* to estimate any subsequent long-run abnormal performance for each firm in this sample. For each event firm, the *BHAR* is the difference between the long-run holding period return for that firm and the long-run holding period return for some benchmark asset. The benchmark return is a proxy for the normal or expected return of the event firm. Conventionally, the benchmark is a reference portfolio (see Barber and Lyon (1997), Lyon, Barber, and Tsai (1999), and Brav (2000)). That is, each event firm is matched to a portfolio of firms with similar attributes, such as size and/or book-to-market. In the case of the control firm approach, the benchmark is a single firm.

Under the null hypothesis of no abnormal return the expected *BHAR* for each event firm, which we label μ , is zero. This follows because each *BHAR* measures the difference between a firm’s long-run return and the long-run return of a proxy designed to have the same expected long-run return. Hence, the *BHAR* measures any abnormal

performance and idiosyncratic risk. The idiosyncratic risk has an expectation of zero and covariance matrix Σ under the null hypothesis.

Our approach can be thought of as consisting of two parts. The first is that we consider multiple *BHARs* per event firm and use a Wald statistic to control for the correlation between observations induced by this approach. The use of multiple measures per event firm can be seen as a way to increase the number of observations and so the power of tests for abnormal performance. Secondly, the use of a (relatively) few control firms per event firm implies that the selection of control firms can be done in such a way as to provide well-specified and powerful tests in a variety of circumstances.

1.3 The Test Statistic

For a sample of N firms, we consider the long-run performance for each firm over an interval of length τ , which is divided into T periods of equal length. In this study, τ equals one year which is divided into $T = 12$ months, each indexed by $t = 1, \dots, 12$.⁶ For each event firm i , we consider the event date to be date zero. Firm i 's gross holding-period return, $r_{i\tau}$, is then the product of the gross monthly returns, $r_{it} = \prod_{t=1}^T r_{it}$. For each event firm, there are M associated control firms, $m=1, \dots, M$, each with a gross holding-period return denoted by $r_{m\tau} = \prod_{t=1}^T r_{mt}$. The *BHAR* for a event firm with respect to a given control firm is defined as $BHAR_{im} \equiv \hat{\mu}_{im} \equiv r_{i\tau} - r_{m\tau}$. The *BHAR* with respect to a reference portfolio p is given by $\hat{\mu}_{ip} \equiv r_{i\tau} - r_{p\tau}$.

⁶ We focus on a annual holding period for two reasons. First, the results are very similar for longer holding periods thus in the interest of space we present results only for a holding period of one year. The techniques are easily extended to account for any desired holding period. Furthermore, typical investors appear to turnover their portfolios frequently so a relatively short holding period is appropriate if one is to measure "investor experience" (Benartzi and Thaler (1995) estimate the typical investor's investment horizon as one year, see also Barber and Odean (2000)).

For each event firm we compute a set of M *BHARs*, one for each control firm. To test hypotheses about linear combinations of *BHARs*, we compute the Wald statistic

$$W = (a - A\hat{\Gamma})'[A(S \otimes I)A']^{-1}(a - A\hat{\Gamma}),$$

where $\hat{\Gamma}' = [\hat{\mu}_{11}, \hat{\mu}_{12}, \dots, \hat{\mu}_{1M}, \dots, \hat{\mu}_{N1}, \hat{\mu}_{N2}, \dots, \hat{\mu}_{NM}]$. That is, $\hat{\Gamma}$ is a column vector of *BHARs* of length NM . The first M rows of $\hat{\Gamma}$ are the *BHARs* of the first event firm with respect to each of its M reference assets. The next M rows in $\hat{\Gamma}$ are the *BHARs* of the second event firm with each of its M reference assets, and so on. The cross-sectional covariance structure of the *BHARs* is represented by the contemporaneous covariance matrix $S_{M \times M}$, where S is our estimate of Σ . The covariance matrix of $\hat{\Gamma}$ has a block-diagonal structure $S \otimes I$, where I is an $N \times N$ identity matrix.

The test of whether the equally weighted average *BHAR* equals zero considers whether $a - A\hat{\Gamma} = 0$, where the scalar $a = 0$ and A is row vector of ones corresponding to the NM rows of the $\hat{\mu}_{im}$ in $\hat{\Gamma}$. Under the null, the test statistic W is drawn from a $\chi^2(1)$ distribution. Value weighted schemes are easily considered by altering the vector A .

In contrast, the standard test of the hypothesis that the expected *BHAR* is zero (using a portfolio as the reference asset) computes the sample average of N *BHARs*, $(\frac{1}{N} \sum_{i=1}^N \hat{\mu}_{ip})$, and tests whether it is statistically significantly different from zero using a t -test. The reason this test is commonly misspecified is that each $\hat{\mu}_{ip}$, the difference between a skewed holding-period return on an individual security and a much-less-skewed portfolio return, is positively skewed. This positive skewness implies that the average *BHAR* is likely to be negative when the variance of the *BHAR* is low, leading to high rates of rejection when the null is true. This is in contrast to any control firm based

approach where the *BHAR* is the difference between two holding-period returns with approximately equal skewness and variance.⁷

Some basic properties of the Wald statistic can be seen directly from its definition. First, as the correlation between control firm returns increases, *ceteris paribus*, the value of the test statistic decreases, making it harder to reject the null. To see this, we first define some notation. Let the sample covariance between *BHARs* defined using control firms j and k be $\hat{\sigma}_{jk}$. Further, when $j=k$, $\hat{\sigma}_{jk} = \hat{\sigma}_j^2$, the variance of the *BHAR* using control firm j . The denominator of our test statistic, $[A(S \otimes I)A']$, is then $\sum_{j=1}^M \sum_{k=1}^M \hat{\sigma}_{jk}$, the sum of the variances of the *BHARs* plus twice the sum of the covariances of the *BHARs* within each block of the covariance matrix. Now, fixing the variances of the individual *BHARs*, and increasing $\hat{\sigma}_{jk}$, $j \neq k$ has the effect of increasing $[A(S \otimes I)A']$ and reducing W . In the limit, when the correlation between control firms equals one (so no new information is added after the first control firm), W becomes the square of the t -statistic obtained when using one control firm per event firm.

Second, the numerator of the test statistic is the average *BHAR* across the M control firms for each event firm and across the N event firms. Therefore, a well specified test requires, for each event firm, the average *BHAR*, across the M control firms, to be zero under the null. It is not necessary that the mean return of event firm i be matched by the mean return of each of its M control firms, rather the mean return of the event firm need only match the average of the mean returns for the M control firms.

⁷ For the difference of two approximately-equally-skewed distributions not to be skewed, it is well-known that the variances of the two distributions must also be approximately equal. Empirically, even without controlling for equivalent variances, *BHAR* skewness under the control firm approach is minimal.

1.4 Control Firm Selection

In simulations using artificial “monthly stock returns” generated using identical (but not independent) lognormal distributions, we found that the optimal way to match event and control firms was to find M control firms that were each highly correlated with the event firm, but which had correlations with each other that were as low as possible. This is equivalent to running a *population* regression of each event firm’s holding-period return on every possible combination of M matching firm returns and choosing the combination of M firms that generates the highest R^2 . In this way, for any given M , the set of control firms add the maximum amount of information, making the test as powerful as possible.⁸ The discussion above concerning the properties of the test statistic illustrates the intuition behind such a procedure.

With actual returns additional issues arise. First, if event firm mean returns are not matched by the average of the control firm mean returns, the size of the test is affected.⁹ Further, if the average of the mean returns of the control firms equals the mean return for each event firm, but each individual control firm does not have the same mean return as the event firm, the added noise will lower the power of the test. Second, because correlations among firm returns are not necessarily the same in the estimation period as in the event period, using a purely statistical procedure to select control firms can result in over-fitting and consequently a test with low power. Some control for this potential problem must therefore be used in any statistically-driven matching procedure.

⁸ In the simulations, control firm and event firm mean returns were the same and inter-firm return correlations were the same across all control firms.

⁹ This statement conveys the intuition but is not precisely true. What is required is that across the observations the expected return of the each reference asset is an unbiased estimate of the expected return of its associated event firm. This is why samples that are biased towards particular firm specific characteristics known to be related to expected returns cause problems for the reference portfolio approach.

Our first approach matches event firms with control firms based on pre-selected firm characteristics. For each event firm we select M control firms that match the event firm as closely as possible on some combination of characteristics which have been found to be related to the cross-section of average returns. Barber and Lyon (1997), for example, show that for random samples of event firms a well specified test is obtained by selecting control firms matched on size and book-to-market. We employ the following matching criteria: size; book/market; size and book/market (Barber and Lyon (1997)); beta; buy-and-hold return momentum (geometric returns); cumulative return momentum (arithmetic returns); industry (2 digit SIC code); industry and size; industry and book/market; industry, size and book/market (as in Barber and Lyon (1997)); industry and beta; industry and buy-and-hold return momentum; industry and cumulative return momentum. We construct and compare tests using different numbers of control firms (different M) with different combinations of the above mentioned matching criteria in order to determine whether different combinations of criteria materially affect the results.

Our second approach we examine is a statistically based “maximal R^2 ” matching procedure. In order to control the possible problems associated with a statistically driven approach we divide this procedure into two parts. First, we reduce the number of potential control firms to K by initially matching on firm characteristics in order to match the mean return of the event firm and the candidate control firms. For example, for each event firm i , one might choose S_z firms that match on size, B firms that match on beta, and BM firms that match on book-to-market such that $S_z + B + BM = K$. Matching occurs in

the event month. In general, we let $K=30$.¹⁰ From this set of K firms, we pick the M that maximize the R^2 of the following regression

$$r_{it} = \alpha + \sum_{j=1}^M \beta_j r_{jt} + \varepsilon_{it}$$

estimated over the 60 months preceding the event month, $t = -60, \dots, -1$.

After specifying the set of K candidate control firms, there are two ways to select the M control firms: unconstrained choice from the entire set of K firms, and constraining the number of firms drawn from each set of control firms (e.g. size, beta, and book-to-market) to be equal. We use the former approach in our simulations. The advantages of the latter are that the number of possible sets of M is greatly reduced and the chance of matching the mean of the event firm is increased. Further, the off-diagonal terms in S , the estimate of the contemporaneous covariance matrix for the *BHAR*'s (Σ) then have a natural interpretation. The first vector of *BHAR*s contains firms matched, for example, by size only; the second contains firms matched by book-to-market, and so on. Each off-diagonal term then represents the cross-sectional covariance between *BHAR*'s matched by different attributes. Matching in this way also has the advantage that the contemporaneous covariances between these *BHAR* vectors may be lower. Empirically, however, we show that there is little difference in power between the two approaches.

2. Data and Simulation Methodology

For the purpose of comparison we match our data set with that of Barber and Lyon (1997) as closely as possible. Our data is monthly returns, closing prices and

¹⁰ The choice of $K = 30$ balances an increase in the power of the test that comes from selecting the control firms from a broader set against the increase in computing time necessary to search over the broader set.

shares outstanding for all NYSE, AMEX, and NASDAQ ordinary common shares available from CRSP between December 1962 and December 1994. The corresponding book value of common equity data is from COMPUSTAT.

We define firm size to be the market value of common equity (shares outstanding multiplied by the monthly closing price). Following Barber and Lyon (1997), we calculate firm size in June of each year (i.e. June shares outstanding times the June closing price) and use it as the observation of size for the subsequent twelve months (from July of the same year to June of the following year) for the purpose of size-based ranking. A firm's book-to-market ratio is the book value of common equity divided by the market value of common equity. As in Barber and Lyon (1997), we measure the book-to-market ratio in December of each year. Rankings based on book-to-market ratios use the December value of the ratio for the observations of the book-to-market ratio for July of the next year through June of the year after.

This study examines only annual *BHARs*. As discussed above, this is a reasonable representation of the average holding period for individual investors. Furthermore, the results for 3 and 5 year horizons are essentially the same. We therefore remove from our sample event firm-months that do not have 12 consecutive subsequent reported monthly returns and those that lack size and book-to-market information. Firm months with book values of common equity less than or equal to zero are also omitted.

In order to generate a standard simulated "event sample" we follow a procedure comparable to that used in Barber and Lyon (1997). For each firm-month in the total sample and for each possible control firm (or reference portfolio) we calculate an annual *BHAR*. That is, for the January 19X1 return on a given pair of sample and control firms

we calculate the annual *BHAR* using the 12 months January – December of 19X1. From the same pair's February 19X1 returns we calculate the annual *BHAR* using the 12 months February 19X1 – January 19X2 and so on. In this way we generate a panel of annual *BHARs*. We randomly draw $L = 10,000$ random samples of $N = 200$ event firm months and for each event firm we calculate its *BHARs* with respect to each of M selected control firms. We then induce an abnormal return for each event month by adding a constant amount between -0.2 and 0.2 to each of the *LNMBHARs*. Test statistics for each of the L samples are calculated and the size and power of the tests are computed.

When comparing with their tests we follow the Barber and Lyon (1997) procedure for building reference portfolios and for selecting control firms. In particular, we match event firms to control firms based on size and then book-to-market ratios the most successful matching procedure used by Barber and Lyon. When we form reference portfolios we follow their procedure to eliminate the new listing and rebalancing biases.

3. Test Results

3.1 Random Samples

3.1.1 Comparison of Test Properties

Here we document the size and power of our test for *BHARs* at a one year horizon for random samples of event firm months. We also compare the size and power of the Wald test to that of the bootstrapped skewness adjusted t-statistic, the empirical p-value, and control firm approaches examined in Lyon, Barber, and Tsai (1999).

In figure 1, we compare the size and power of the different tests considered by Lyon, Barber, and Tsai (1999) and our Wald test using both a maximal R^2 approach and a

characteristic approach. Our rejection rates are based on 10,000 random samples of 200 event months (the Lyon, Barber and Tsai simulations are based on 1,000 random samples). Several results are noteworthy. First, as expected, all of the tests have the appropriate size. Second, as reported by Lyon, Barber and Tsai (1999), the power of the t-statistic using the control firm approach is lower but more symmetric than the power of the empirical p-value or the skewness adjusted t-statistics using the reference portfolio approach.¹¹ Third, the Wald test with the maximal R^2 approach to selecting 3 control firms is the most powerful of all the tests at all levels of induced abnormal returns. Finally, the most powerful Wald test using a characteristic approach is more powerful than the empirical p-value or the skewness adjusted t-statistics for positive induced abnormal returns but, due to the skewness in the reference portfolio approaches, slightly less powerful than the empirical p-value approach for negative induced abnormal returns.

Figure 1 illustrates that with our maximal R^2 approach, using only 3 control firms per event firm, the benefits of the control firm approach remain while the power of the test increases dramatically. In essence, the benefit of the reference portfolio approach (increased information) is gained without introducing the skewness that comes (via diversification) with the use of a portfolio as a control asset. The messages of figure 1 are confirmed in table 1a which provides numerical documentation of the size and power of the relevant test statistics.

¹¹ The asymmetry in the power of the test statistics developed by Lyon, Barber, and Tsai (1999, figure 1) using a reference portfolio approach is due to *BHAR* skewness. With a positively skewed distribution for the *BHAR*, adding a small negative abnormal return to a mean zero *BHAR* increases the probability of rejecting the null very rapidly while the addition of a small positive abnormal return increases the probability of rejecting the null relatively slowly. Further, adding sufficiently large positive abnormal returns will guarantee rejection of the null (given the truncation of the left side of the distribution). Addition of large negative abnormal returns may not guarantee rejection of the null given the large right tail of the distribution under the null. Although Lyon, Barber and Tsai's refinements yield well-specified tests, the asymmetry in the test's power remains as a byproduct of the skewness in the *BHARs*.

As a robustness check table 1b presents the size and power for a representative set of tests using the characteristic based matching approach in the Wald test. For illustration we present results from matching on a variety of firm characteristics that have been shown to explain the cross-section of average returns. There is surprisingly little variation in the size or power of the tests developed based on different sets of characteristics. The table presents “average” rather than extreme results for each value of M . We note only that the test has the correct size for all values of M .

Although our method of search for control firms was guided by theory, our procedure is in no sense optimal. For one thing, we made no attempt to force the average $BHAR$ to zero, which could be done using the coefficient estimates $\hat{\alpha}_i$ and $\hat{\beta}_{im}$ from an estimation period regression of each event firm’s return on the return of each control firm (m) ($r_{it} = \alpha_i + \beta_{im} r_{imt} + \varepsilon_{imt}$). Then, in the event period, the $BHAR$ for a particular event firm-control firm pair would be given by: $BHAR_{im} = r_{it} - \hat{\alpha}_i - \hat{\beta}_{im} r_{mt}$. Such a procedure is likely to generate problems however since average estimates of $\hat{\beta}_{im}$ are less than one which would reintroduce the skewness problem. Empirical comparative static analyses on other aspects of the two matching procedures are considered next.

3.1.2 Empirical Comparative Statics on M

The Wald test using the maximal R^2 approach and 3 control firms per event firm provides a more powerful test than the others examined in figure 1. An important question concerns the choice of M , the number of control firms selected for each event firm. Figure 1 uses $M = 3$ control firms for each event firm in the maximal R^2 procedure. *A priori*, there is no obvious reason for this choice. In table 2, we compare the size and power of the Wald test using the maximal R^2 approach for different values of M : 1, 2, 3,

and 4. The table shows that across the different levels of induced abnormal return there is an almost monotonic increase in power with M , up to $M=3$. Further increases in M lead to marginal increases in power at best. This, as well as the increase in computing power and time required by larger choices of M leads us to choose $M = 3$ as our benchmark for this control firm selection procedure. Note however that the set of potential control firms includes only firms that match on size, book-to-market, and beta. The use of other (or multiple) initial matching criteria may suggest a different choice.

In table 3 we examine a variation of the analysis of table 2 to determine if we can increase power by better matching the expected return of each control firm more closely to that of the event firm. We predict that this will increase the power of the tests for a given M as the volatility of the individual *BHARs* should be reduced. To accomplish this, we use an initial matching criterion that selects candidate firms matched on two firm characteristics, (rather than one as in table 2). We again select $K = 30$ candidate control firms where 10 are selected that match the event firm most closely based on firm size and then on book-to-market ratio, 10 are selected that match on size and then on beta, and 10 are selected that match most closely on book-to-market and then beta.

Here we again see that setting $M = 3$ is a reasonable choice. Power increases with M until $M=3$. Increasing M to 4 again results in a minor improvement. Finally, by comparing table 1a to table 3, we see that the use of two characteristics in the initial matching criteria does not bring meaningful increases in power. The differences between the results of table 1a for the maximal R^2 procedure and those reported in table 3 for $M = 3$ are not statistically significant. The absence of the predicted improvement may be the

result of sampling error. Repeating simulations of 10,000 random draws for the same conditions has, however, shown little variation in the aggregate outcomes.

Empirical comparative static checks on the appropriate value for M in the characteristic-based matching procedure are also possible. Note however that in the characteristic-based procedure it is difficult to vary only M in these comparisons. We have performed a large variety of checks and find that they all support the conclusion that the use of three control firms per event firm is a reasonable choice. To facilitate comparison with the control firm approach of Barber and Lyon (1997) we present a comparison of versions of the characteristic-based selection procedure with $M = 1, 2, 3, 4,$ and $5,$ in which all the control firms are matched based first on size and then book-to-market ratio. Table 4 presents these tests. Consistent with the findings of Barber and Lyon (1997) we see that the Wald test has the correct size for all choices of M ; empirical rejection levels when there is no induced abnormal return are very close to theoretical levels, for each of the standard significance levels. Finally, the power of the test, at all levels of induced abnormal return, increases monotonically with M . The largest increases in power occur as M increases from 1 to 3. Increasing M further once again produces only marginal increases in power for all levels of induced abnormal return.

3.1.3 Other Empirical Comparative Statics

In table 5, we compare the results of using the maximal R^2 procedure when the search is constrained to select one firm from each set of attributes used in the initial matching process versus when the search is an unconstrained maximization of the R^2 . As the table shows, such a constraint is not costly in terms of power. This is not surprising. Given that all firms associated with a given attribute were selected from the closest

matches for that single attribute, the correlation between such firms will tend to be higher. Thus, the matches selected using unconstrained maximization should tend to be firms matched on different attributes.

Finally, given the success of the maximal R^2 procedure, it is appropriate to ask whether the improvement in power is due to better control firm selection, or whether it is due to the use of the Wald test itself. In order to address this question, we use the empirical p-value approach of Lyon, Barber, and Tsai (1999), but use the maximal R^2 procedure to construct a reference portfolio of ($M =$) 3 firms. That is, we use the weights from the estimation period (maximal R^2) regression to create the returns on a reference portfolio for event firm i : $r_{ip} = \prod_{t=1}^{12} (\hat{\alpha} + \sum_{j=1}^3 \hat{\beta}_j r_{jt})$. Then, for each event firm, the $BHAR_{ip}$ is $r_{i12} - r_{ip}$. To correct for the resulting skewness of this $BHAR$, we employ the empirical p-value technique of Lyon, Barber, and Tsai (1999). The results are given in table 6. Columns 2 and 3 in the table present Lyon, Barber, and Tsai's (1999) results using the empirical p-value and our empirical p-value tests using the maximal R^2 procedure for reference portfolio selection. Column 4 presents the results of our maximal R^2 procedure. Observe that our test procedure itself contributes most to improvements in power. Indeed, it is not clear whether using the maximal R^2 procedure to form a reference portfolio represents an improvement on Lyon, Barber, and Tsai's (1999) approach. The power is higher for positive induced abnormal returns, but lower when induced abnormal returns are negative. This is due to the difference in skewness of $BHARs$ created with standard reference portfolios versus a reference portfolio of only 3 firms.¹²

¹² A similar comparison for the maximal R^2 procedure and the control firm approach can be found in tables 1a and 3. Results for the use of a t-statistic and the selection of a control firm via a maximal R^2 procedure

3.2 Nonrandom Samples

That the Wald statistic approach provides well specified tests in random samples of event firms is not surprising in that it corrects for the rebalancing, new issue, and skewness biases. However, when the sample of event firms is not random the “bad model” problem and the concern over non-independent observations are heightened. The extent to which these issues affect the size of the Wald test is an empirical question that is examined below. However there are reasons to expect that the approach introduced here will provide a well specified test.

If expected returns are related to observable firm characteristics then to the extent that the event sample is biased towards one (or more) such characteristic(s) and the test for abnormal performance does not control for the implied change in normal performance of event firms, the test will be misspecified. In this subsection we demonstrate that the flexibility imparted by choosing only a few control firms per event firm allows our matching procedure to control for bias in the sample of event firms and so provide a well specified test. Similarly, if the nature of the sample of event firms causes observations of event firm returns to be correlated (time clustering) the same modification to the matching procedure will allow the Wald test to provide a well specified test.

The impact of the bad model problem in non-random samples is illustrated nicely by the discussion in Lyon, Barber and Tsai (1999) concerning samples of firms with high six month pre-event return performance. They note, for example, that using a control firm approach, matching on size and book-to-market, in a sample of firms with high pre-event return performance, the test statistics are positively biased at the one-year horizon

is found by examining $M = 1$ in table 3. Comparing these results with the results of the control firm approach in table 1a we see that the statistical selection of one control firm provides an increase in power. It is also clear that the use of multiple control firms per event firm provides meaningful increases in power.

but negatively biased at the three- and five-year horizon. The biases are due to the momentum (one-year horizon) and reversal (three-year and five-year horizons) effects documented by Jegadeesh and Titman (1993) and the impact these effects have on subsequent average returns of event firms that are not accounted for by a simple size and book-to-market matched reference portfolio or similarly matched control firms.

The importance of accounting for returns that are correlated across events in the test statistic is illustrated by the discussion in Lyon, Barber, and Tsai (1999) concerning industry concentration. When events are concentrated in a single industry the size of their tests are understated, indicating that the measured variation is too small. This result suggests the presence of a common component of idiosyncratic returns within industries that must be accounted for in order to provide a well specified test. Lyon, Barber, and Tsai (1999) however note that when the event is clustered within four industries, their tests have the correct size. Therefore, empirically, it seems that only extreme industry concentration is a concern. As shown below, when events are clustered in time, it is very important to account for returns that are correlated across events.

In table 7, we report the size of our test for six different biased samples of 200 event firms (each replicated 10,000 times). The samples we consider are taken from subsets of small and large firms, high and low book-to-market firms, and high and low six-month pre-event return performance. Panel A of table 7 shows that a simple characteristic based matching procedure, matching one control firm on size, one on book-to-market and one on beta provides a misspecified test. When the event firms are drawn from a biased sample and that bias concerns a characteristic related to average returns a simple matching procedure fails to match the mean of the event firms with its set of

control firms. Except for the sample of firms with low pre-event returns which is well specified using the standard procedure the other tests generally reject to frequently in the upper tail and to infrequently in the lower tail.

Panel B of table 7 reports the size of tests using a “conditional matching procedure” designed to do a better job of matching the mean return of the event firms with the mean of their control firms and for controlling for common, industry-specific components of idiosyncratic returns. Consider the sample of event firms drawn from the set of small firms. The problem with the test reported in panel A of table 7 is that while one of the matching characteristics is firm size, the other two ignore firm size in their selections of control firms. Control firms that match closely on the book-to-market ratio or the beta of the event firm are likely to differ in the size dimension. Since firm size is known to be related to average return, the expected returns for this set of control firms is unlikely to match the expected return of the event firm resulting in a misspecified test.

The conditional matching procedure employed in panel B of table 7 matches control firms to event firms by taking control firms solely from the subset of firms from which the event firms are drawn. Specifically, for the sample made up of small firms we match control firms to event firms by matching on three characteristics. The first control firm is matched on size alone. The second control firm is matched on firm size and within a set of size matched firms we select the firm with the book-to-market ratio that most closely matches the event firm. The third control firm is also matched first on size and then on industry (2-digit SIC code). The industry based matching of control firms is introduced to control for the presence of any industry-specific component of returns.

Similarly for the samples of firms drawn from biased sets of book-to-market ratios or pre-event return the conditional matching procedure matches on the standard firm characteristics (size, book-to-market, industry) conditional on the potential matches being within the subset of firms from which the event firms are drawn. When the event firms are drawn from the highest or lowest deciles of book-to-market ratio, the three control firms for each event firm are matched on book-to-market, book-to-market and size, and book-to-market and industry. For the event firms drawn from the highest or lowest deciles of six-month pre-event return (momentum), the three control firms match on momentum and size, momentum and book-to-market, and momentum and industry.

As shown in the panel B of table 7, the tests using the conditional matching procedure are all well specified. This emphasizes the importance of matching the mean of the reference asset(s) to the mean of the event firms in long-run event studies and illustrates one benefit of the flexibility of our approach. The results also indicate that Lyon, Barber and Tsai's conjecture that their test would be well specified on samples of firms drawn from the extreme momentum deciles, if they restricted their reference portfolio to contain assets from these same momentum deciles, is likely to be true.

3.3 Calendar Clustering

A further problem discussed by Lyon, Barber, and Tsai (1999) is calendar clustering. Events are often clustered in calendar time. The reason this clustering may lead to measurement problems is that contemporaneous returns are likely to be more highly correlated across firms than non-contemporaneous returns. Mitchell and Stafford (2000) show that when the *BHAR* is calculated using a benchmark that is a portfolio of stocks matched independently on size and book-to-market quintiles they find in

simulations that a lack of independence across *BHARs* is so severe that the true size of a 5% test can be up to 20%! Based on this result, they advocate the use of the calendar-time portfolio approach (see also Fama (1998)) for measuring long-run abnormal returns.

With a correct model for mean returns, however, plus the fact that in such a model, the errors are *i.i.d* (e.g. a factor model such as the APT) calendar clustering would *not* be a problem. As Fama (1998) and Mitchell and Stafford (2000) note, problems can arise if the true asset pricing model permits cross-sectional dependence of the errors (e.g. the CAPM). To determine whether time clustering presents problems for our estimation procedure, in Table 8, we consider the size of the Wald test (using the characteristics-based matching procedure) under the extreme assumption that all “events” occur on the same day – an assumption that should produce maximal cross-sectional correlation in the *BHARs*. The Wald test continues to be well-specified. In Table 9, we examine the size of the Wald test using the maximal R^2 matching procedure when there is calendar clustering. For all choices of $M > 2$ the tests are also well-specified.¹³

These results indicate that under our approach the resulting *BHARs* are only minimally correlated in the cross-section. Both matching procedures appear to do a sufficiently good job of capturing the systematic component of returns. If this accounts for the contemporaneous cross-sectional correlation it is not surprising that calendar clustering does not present a problem for our approach. Furthermore, the use of control firms also implies that the *BHARs* contain the difference between the idiosyncratic returns on two individual assets. Thus any cross-sectional correlation in the idiosyncratic returns of individual assets may not be passed through to the *BHARs* as it would be using

¹³ Note that because the question here is one of size and not power we have set $K = 15$ rather than 30 as in other analysis of the maximal R^2 procedure. This is simply for the savings in computing time required.

a reference portfolio approach. Empirically, we find that the average cross-sectional correlation for the *BHARs* for our time clustered simulations is only .00015. This compares to the average reported in Mitchell and Stafford (2000) of .002, demonstrating the dramatic difference between the reference portfolio approach and our procedure. Finally, this procedure avoids the concern with the calendar time approach noted by Loughran and Ritter (2000) that when calculating average abnormal returns, all months are equally weighted, regardless of the number of observations in a given month.¹⁴

4. Conclusion

In this paper, we propose a simple but powerful test of abnormal long-run holding period returns using the buy and hold abnormal return as the measure of performance. The test is based on a Wald statistic and combines the benefits of the control firm approach (avoids the rebalancing, new listing, and the skewness biases examined by Barber and Lyon (1997)), and the reference portfolio approach (increased power of the test due to the reduction of the variance of the measure of abnormal return).

We can think of the innovation in our approach in two different ways. First it can be envisioned as considering the difference in the returns of an event firm's equity and the equity of each of the firms in a reference portfolio. The result is a set of *BHARs* for each event firm where the distribution of each *BHAR* is no longer skewed. Forming an "equally weighted portfolio" of such *BHARs* then results in a test with high power. Rather than forming a *BHAR* for each event firm by taking the difference between its

¹⁴ They also point out that the problem is not just for calculating statistical significance but also of potentially missing an endogenous response. For example, if firms issue seasoned equity when their stock prices are currently high and mispricing is correlated across securities, so that seasoned equity offerings are too, the calendar time approach will miss this endogenous response. Our conditional matching procedure can be used to address this issue.

long-run return and that of a portfolio we form a set of *BHARs* for each event firm and then combine them to increase the power of the test. Secondly, we can think of the Wald test as being a standard t-test on the sample of *NM* observations where the variance is adjusted to account for the induced cross-sectional correlation between observations. The use of multiple control firms for each event firm is then a way to increase the sample size and so the power of the test without the need to increase the number of event firms.

The Wald test is shown to be well-specified in random as well as non-random samples of event firms. The key insight is that when choosing control firms, it is important to select them from the subset of firms from which the sample was drawn. The flexibility to accomplish this vital matching is one of the benefits of our approach. Our procedure also provides a well-specified test statistic even in simulations where events are maximally clustered in time. Even in this extreme circumstance our matching procedures generate *BHARs* that are only minimally correlated across event firms.

If the buy and hold abnormal return is the variable of interest in a test for the absence of long-run abnormal performance against a specific alternative hypothesis, the approach developed in this paper appears to alleviate the statistical concerns associated with using buy and hold abnormal returns for evaluating abnormal performance. This is important because in such a circumstance, the finding of an absence of abnormal performance based upon long-run cumulative abnormal return does not imply the absence of long-run buy and hold abnormal performance. However, if the alternative hypothesis is simply a vaguely defined notion of market inefficiency, then either performance metric may be seen to capture the presence or absence of market inefficiencies. A conservative approach to long-run event studies, which seems prudent given the difficulty associated

with specifying long-run normal returns, would be to consider both a calendar time approach utilizing average periodic abnormal returns and the approach developed here based on the long-run buy and hold abnormal returns.

An issue of ongoing study is the most appropriate selection of control firms. We have introduced two approaches, a maximal R^2 approach and a characteristic approach, both of which serve to greatly increase the power of the test relative to the standard t-statistic approach using a single control firm for each event firm. The maximal R^2 approach examined here is well-specified and has power that is greater than any of the tests proposed by Lyon, Barber, and Tsai (1999) but may not be the most powerful approach available using this framework.

References

Barber, B.M., and J.D. Lyon, 1997, "Detecting Long-Run Abnormal Stock Returns: The Empirical Power and Specification of Test Statistics," *Journal of Financial Economics* 43, 341-372.

Barber, B.M., and T. Odean, 2000, "Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors," *Journal of Finance* 55, 773 – 806.

Benartzi, S. and R.H. Thaler, 1995, "Myopic Loss Aversion and the Equity Premium Puzzle," *Quarterly Journal of Economics* 110, 73 – 92.

Bernard, V. L., 1987, "Cross-sectional Dependence and Problems in Inference in Market-based Accounting Research," *Journal of Accounting Research* 25, 1 – 48.

Brav, A., 2000, "Inference in Long-Horizon Event Studies: A Bayesian Approach with Application to Initial Public Offerings," *Journal of Finance* 55, 1979-2016.

Canina, L., R. Michaely, R. Thaler, and K. Womack, 1996, "A Warning About Using the Daily CRSP Equally-Weighted Index to Compute Long-Run Excess Return," *Journal of Finance* 53, 403-416.

Collins, D. W. and W. T. Dent, 1984, "A Comparison of Alternative Testing Methodologies Used in Capital Market Research," *Journal of Accounting Research* 22, 48 – 84.

Fama, E.F., 1998, "Market Efficiency, Long-Term Returns, and Behavioral Finance," *Journal of Financial Economics*, 49, 283-306.

Fama, E.F., and K. French, 1993, "Common Risk Factors in Returns on Stocks and Bonds," *Journal of Financial Economics* 33, 3056.

Jegadeesh, N. and S. Titman, 1993, "Returns to Buying Winners and Selling Losers: Implications for stock Market Efficiency," *Journal of Finance* 48, 65-91.

Kothari, S. P., and J. B. Warner, 1997, Measuring Long-Horizon Security Price Performance, *Journal of Financial Economics* 43, 301-340.

Lyon, J.D., B.M. Barber, and C. Tsai, 1999, "Improved Methods for Tests of Long-Run Abnormal Stock Returns," *Journal of Finance*, 54, 165-201.

Loughran, T. and J. Ritter, 1995, "The New Issues Puzzle," *Journal of Finance* 50, 23 – 51.

Loughran, T., and J. Ritter, 2000, "Uniformly Least Powerful Tests of Market Efficiency," *Journal of Financial Economics*, 55, 361-389.

Mitchell. M., and E. Stafford, 2000, "Managerial Decisions and Long-Term Stock Price Performance," *Journal of Business*, 73, 287-329.

Ritter, J., 1991, "The Long-Term Performance of Initial Public Offerings," *Journal of Finance* 46, 3 – 27.

Sefcik, S. E. and R. Thompson, 1986, "An Approach to Statistical Inference in Cross-sectional Models with Security Abnormal Returns as Dependent Variables," *Journal of Accounting Research* 24, 316 – 334.

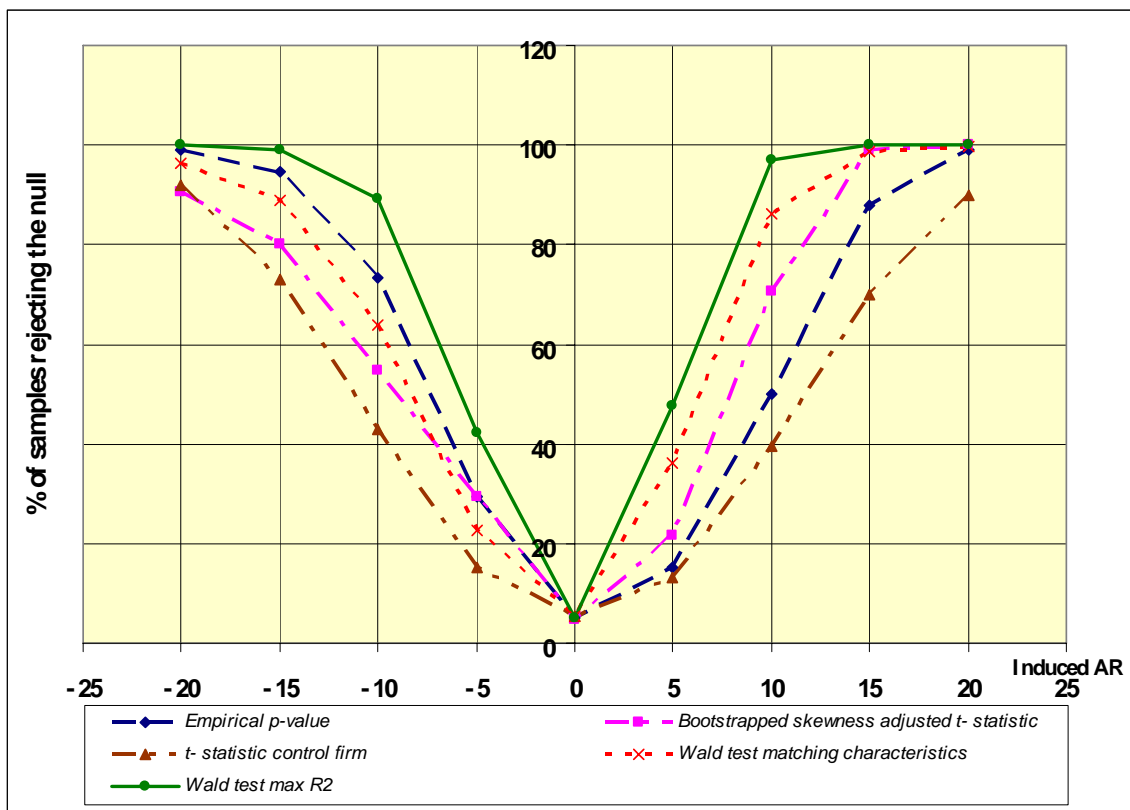


Figure 1. Comparison of the power of test statistics in random samples. In each simulation, there are 200 event firms. For our tests, we run 10,000 simulations. The control firm with standard t-statistic, bootstrapped skewness-adjusted t-statistic and empirical p-value results come from Lyon, Barber, and Tsai (1999), who run 1000 simulations. For the maximal R^2 procedure, the number of potential matching firms, $K=30$, so that $S_z=B=BM=10$. The number of control firms for each event firm, M , is 3. The characteristic-based matching procedure, which represents our best result for $M=6$, uses the following matching characteristics: size; book-to-market; SIC and size; SIC and book-to-market; SIC and beta; SIC, size, and book-to-market. The horizontal axis in the figure is the induced abnormal return. The graphs represent the percentage of the time that the null hypothesis of no annual buy-and-hold abnormal return, is rejected.

Induced annual abnormal returns	Bootstrapped Skewness adjusted t- statistic	Empirical p value	t- statistic control firm	Wald test Max R ²	Wald test characteristic approach
-20	90.7	99.1	92.0	100.0	96.4
-15	80.1	94.7	73.0	99.2	89.0
-10	54.8	73.4	42.8	89.4	64.0
-5	29.3	29.3	15.1	42.1	22.5
0	4.9	5.1	5.3	5.0	5.4
5	21.7	15.1	13.3	47.5	36.3
10	70.5	50.1	39.5	96.9	86.1
15	99.1	88.0	70.1	100.0	98.8
20	100.0	99.1	89.8	100.0	99.8

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 1a. Figure 1. Comparison of the power of test statistics in random samples. These numbers are used to create figure 1.

M	3	4	5	6
Induced annual abnormal returns	Size, BtM, Beta	Size, BtM, SIC&Size, Beta	Size, BtM, Size&BtM, SIC&Pi Momentum, SIC&CARMomentum	Size, BtM, SIC&size, SIC&BtM, SIC&Beta, SIC&Size&BtM
-20	95.5	96.3	96.2	96.4
-15	85.1	88.2	88.6	89.0
-10	56.1	61.1	65.7	64.0
-5	18.6	20.5	26.6	22.5
0	5.1	5	5.3	5.4
5	27.0	32.9	19.3	36.3
10	74.8	81.6	67.2	86.0
15	96.0	98.0	95.1	98.8
20	99.4	99.6	99.6	99.8

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 1b. Comparison of the power of test statistics in random samples of characteristic-based matching for different numbers of matching firms and different characteristics. In each simulation, there are 200 event firms. For out tests, we run 10,000 simulations. The numbers in the table represent “typical” results (with the exception of $M=6$, which is the “best” result shown in table 1a) in the sense that for each number of matching firms the use of alternative characteristics yields very similar results.

		M			
Induced annual abnormal returns		4	3	2	1
Power	-0.2	98.8	100.0	98.8	98.6
	-0.15	97	99.2	95.5	94.3
	-0.1	80.4	89.4	79.9	74.9
	-0.05	39.5	42.1	37.1	31.9
	0	4.5	5.0	5.9	5.8
	0.05	26.1	47.5	25.9	21.4
	0.1	84.9	96.9	79.8	66.6
	0.15	99.4	100.0	98.0	93.5
	0.2	100	100.0	99.7	98.9
Size	0.5	0.5	0.4	0.5	0.5
	99.5	0.7	0.3	0.7	0.5
	2.5	2.3	2.6	2.4	2.2
	97.5	2.1	2.1	3.0	2.7
	5	4.5	5.1	4.7	4.3
	95	4.5	4.5	5.9	5.8

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 2 Effect of number of matching firms on size and power of Wald tests using maximal R² matching. The number of potential matching firms, $K=30$, so that $Sz=B=BM=10$. In each simulation, there are 200 event firms. We run 10,000 simulations. For the power comparison, the numbers in column 2 are the induced abnormal returns. For the size comparison, the induced abnormal return is zero; the numbers in column 2 represent the theoretical size of the two-tailed test. The numbers in the rest of the table represent the percentage of the time that the null hypothesis that there is no annual buy-and-hold abnormal return, is rejected.

		M			
Induced annual abnormal returns		4	3	2	1
Power	-0.2	99	99.2	98.9	98.8
	-0.15	95.5	97.6	95.7	94.3
	-0.1	81.7	81.3	79.6	73.2
	-0.05	38.9	39.1	36.1	29.2
	0	5.0	4.8	5.3	4.9
	0.05	28.6	30.9	29.4	24.4
	0.1	88.0	85.5	84.0	71.0
	0.15	99.8	99.7	98.9	95.0
	0.2	100	100	99.8	99.1
Size	0.5	0.8	0.7	0.5	0.7
	99.5	0.8	0.5	0.5	0.4
	2.5	2.6	2.3	2.3	2.5
	97.5	2.4	2.4	2.7	2.3
	5	4.9	4.7	4.6	4.6
	95	5.0	4.8	5.3	4.9

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 3. Effect of number of matching firms on size and power of Wald tests using maximal R^2 matching. The number of potential matching firms, $K=30$, so that $S_{z\&B} = S_{z\&BM} = BM\&B = 10$. In each simulation, there are 200 event firms. We run 10,000 simulations. For the power comparison, the numbers in column 2 are the induced abnormal returns. For the size comparison, the induced abnormal return is zero; the numbers in column 2 represent the theoretical size of the two-tailed test. The numbers in the rest of the table represent the percentage of the time that the null hypothesis that there is no annual buy-and-hold abnormal return, is rejected.

	Induced annual abnormal returns	M				
		1	2	3	4	5
Power	-0.2	92.1	94.8	95.6	96.0	96.3
	-0.15	76.0	83.5	86.0	87.4	88.2
	-0.1	45.5	55.2	60.0	62.3	64.1
	-0.05	15.7	20.6	23.2	25.1	26.5
	0	4.7	4.8	5.2	5.3	5.5
	0.05	15.3	17.3	18.3	18.7	18.4
	0.1	44.9	55.6	60.2	63.2	64.6
	0.15	75.7	87.2	91.5	93.4	94.5
	0.2	91.5	97.3	98.9	99.4	99.5
Size	0.5	0.4	0.4	0.4	0.5	0.5
	99.5	0.4	0.6	0.6	0.7	0.8
	2.5	2.3	2.1	2.1	2.4	2.5
	97.5	2.4	2.5	2.6	2.8	2.8
	5	4.5	4.5	4.7	4.7	4.7
	95	4.7	4.8	5.2	5.3	5.5

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 4. Comparison of the power of test statistics in random samples for characteristic-based matching for different numbers of matching firms. Here, we standardize the characteristic-based procedure. For all M , we first match on size decile. We then find the M closest match(es) by book-to-market. In each simulation, there are 200 event firms. For our tests, we run 10,000 simulations. The numbers in the table for represent the percentage of the time that the null hypothesis that there is no annual buy-and-hold abnormal return, is rejected.

	Induced annual abnormal returns	Unconstrained matching	Constrained matching
Power	-0.2	100.0	100.0
	-0.15	99.2	99.0
	-0.1	89.2	88.0
	-0.05	40.7	40.4
	0	5.2	4.9
	0.05	46.7	45.9
	0.1	96.9	96.3
	0.15	100.0	100.0
	0.2	100.0	100.0
Size	0.5	0.6	0.4
	99.5	0.6	0.5
	2.5	2.7	2.3
	97.5	2.8	2.4
	5	5.0	4.8
	95	5.2	4.9

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 5. Restrictions on the maximal R^2 procedure. We compare the size and power of our tests when restricting our procedure to select one matching firm each characteristic with the size and power when the restriction is not imposed. The number of potential matching firms, $K=30$, so that $S_z=B=BM=10$. In each simulation, there are 200 event firms, and the number of matching firms for each event firm, $M=3$. We run 10,000 simulations. For the power comparison, the numbers in column 2 are the induced abnormal returns. For the size comparison, the induced abnormal return is zero; the numbers in column 2 represent the theoretical size of the two-tailed test. The numbers in columns 3 and 4 represent the percentage of the time that the null hypothesis that there is no annual buy-and-hold abnormal return, is rejected.

	Induced annual abnormal returns	LBT (1999) Empirical p value	Reference Portfolio of 3 firms generated using maximal R ² procedure	Maximal R ² procedure with 3 matching firms
Power	-0.2	99.1	99.3	99.6
	-0.15	94.7	94.6	97.8
	-0.1	73.4	69.2	83.3
	-0.05	29.3	23.3	38.7
	0	5.1	5.0	5.2
	0.05	15.1	16.8	23.7
	0.1	50.1	58.8	83.4
	0.15	88.0	93.3	99.5
	0.2	99.1	99.4	100.0
Size	0.5	0.4	0.61	0.2
	99.5	0.9	0.58	0.7
	2.5	2.6	2.45	3.2
	97.5	2.5	2.54	2.5
	5	5.0	4.67	6.2
	95	4.8	5.16	5.2

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 6. Size and power using maximal R² matching with Wald tests and pseudoportfolio procedure.

The number of potential matching firms, $K=15$, so that $S_z=B=BM=5$. In each simulation, there are 200 event firms, and the number of matching firms for each event firm, $M=3$. We run 10,000 simulations.

For the pseudoportfolio procedure we use 10,000 pseudoportfolios that each contains firms randomly drawn from the same size/book-to-market portfolio as the event firm, to generate the empirical distribution and find the p-value corresponding to the chosen significance level.

For the power comparison, the numbers in column 1 are the induced abnormal returns. For the size comparison, the induced abnormal return is zero; the numbers in column 1 represent the theoretical size of the two-tailed test. The numbers in the rest of the table represent the percentage of the time that the null hypothesis that there is no annual buy-and-hold abnormal return, is rejected. Lyon, Barber and Tsai (1999) do not provide the size comparison.

Sample	Matching by	1%			5%			10%		
		0.5%	99.5%		2.5%	97.5%		5.0%	95.0%	
Panel A: matching one control firm on size alone one on book-to-market alone and one on beta alone										
Samples of small firms		0.0*	8.7*	0.5*	27*	0.8*	41.1*			
Samples of firms with low BtM	Size, BtM, Beta	0.4	2.1*	1.8*	6.2*	3.3*	10.6*			
Samples of firms w/ low 6-months pre-event return	Size, BtM, Beta	0.7*	0.3	2.4	2.3	4.5	5.0			
Samples of large firms		0.2*	1.2*	1.9*	3.8*	4.2*	7.3*			
Samples of firms w/ high BtM	Size, BtM, Beta	0.4	2.5*	1.4*	11.2*	2.7*	20.5*			
Samples of firms w/ high 6-months pre-event return	Size, BtM, Beta	0.8*	0.4	1.8*	3.3*	3.8*	6.1*			
Panel B: conditional matching procedure ^a										
Samples of small firms		0.4	0.4	2.5	2.3	5.0	4.9			
Samples of firms w/ low BtM	Size, Size decile & BtM, Size & SIC	0.4	0.5	2.4	2.6	4.7	5.3			
Samples of firms w/ low 6-months pre-event return	BtM, BtM decile & Size, BtM & SIC	0.5	0.6	2.7	2.2	5.3	5.1			
Samples of large firms	Momentum decile & Size, Momentum decile & BtM, Momentum & SIC	0.4	0.4	2.3	2.6	5.2	4.9			
Samples of firms w/ high BtM	Size, Size decile & BtM, Size & SIC	0.4	0.5	2.0	2.5	4.5	4.9			
Samples of firms w/ high 6-months pre-event return	BtM, BtM decile & Size, BtM & SIC	0.6	0.7*	2.3	2.4	4.6	5.1			

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 6. Size with characteristic-based matching in non random samples. In each simulation, there are 200 sample firms. For our tests, we run 10,000 simulations. The induced abnormal return is zero. In column 1, we describe the non-random samples considered. Large (small) firms are in the top (bottom) decile in market value. High (low) book-to-market firms are in the top (bottom) decile in book-to-market. high (low) momentum firms are in the top (bottom) decile in six-month pre-event return. The numbers in the table represent the percentage of the time that the null hypothesis that there is no annual buy-and-hold abnormal return, is rejected.

Matching by	Two-Tailed Theoretical Significance level					
	1%		5%		10%	
	0.5%	99.5%	2.5%	97.5%	5.0%	95.0%
SIC&Size, SIC&BtM, SIC&Size&BtM	0.6	0.7	2.6	2.9	4.9	5.5
Size, BtM, Size&BtM	0.4	0.7	2.5	2.9	4.8	5.6
3 closest from Size&BtM	0.5	0.8	2.5	2.9	5.0	5.4

* Significantly different from the theoretical significance level at the 1 percent level, one sided binomial test statistic.

Table 8. Calendar clustering: size using characteristic-based matching. Column 1 details the three matching criteria. When we match on multiple criteria, the procedure is to first divide into deciles based on size and then find the closest matches on book-to-market within the size decile. When first matching on industry, we match by 2-digit SIC code. In each simulation, there are 200 event firms. All events in a single simulation begin in the same month. We run 10,000 simulations. The headings represent theoretical sizes for our two-tailed tests. The numbers in the table represent actual rejection percentages when there is no induced abnormal return.

Number of matching firms, M	Two-Tailed Theoretical Significance level					
	1%		5%		10%	
	0.5%	99.5%	2.5%	97.5%	5.0%	95.0%
1	0.4	3.5*	1.7	10.5*	2.9*	14.6*
2	0.6	1.2*	2.9	5.3*	5.5	8.0*
3	0.9	0.4	3.0	2.4	4.8	4.9
4	0.7	0.4	2.9	2.4	5.1	5.0
5	0.5	0.5	2.2	2.4	5.4	5.0

* Significantly different from the theoretical significant level at the 1 percent level, one sided binomial test statistic.

Table 9. Calendar clustering: size using maximal R^2 matching as a function of the number of matching firms, M. The number of potential matching firms, $K=15$, so that $S_z=B=BM=5$. In each simulation, there are 200 event firms. All events in a single simulation begin in the same month. We run 10,000 simulations. Column 1 represents the number of matching firms for each simulation. The headings represent theoretical sizes for our two-tailed tests. The numbers in the table represent actual rejection percentages when there is no induced abnormal return.