Abstract

We investigate the use of modern statistical techniques in the application of event studies conducted on single securities for the purpose of securities litigation. Single-firm event studies are widely used in civil litigation, with billions of dollars in settlements hinging on the outcome of the exercise. Prior work has explored modifying the standard single-firm event study design to provide more robust statistical inference. But little work has been done to determine methods that can directly increase the precision of the excess return estimate. We take a prediction approach to the excess return calculation and find substantial performance improvement is possible using modern machine learning methods.
1. Introduction

The event study is one of the most frequently used tools employed by empirical economists in testing the observable impact of events. Widely used by researchers in finance, accounting, and the law, event studies ostensibly allow for the ability to isolate the impact of a broad range of corporate events. They have provided evidence on the consequences of legal and regulatory changes, the proposed benefits and costs of mergers, and the implications of corporate takeover policies. Event studies have also featured prominently in the decades-long American experiment with private securities litigation.

The event study technique was first used in the 1960s by financial economists to test the speed of adjustment of prices to new information, in particular to the announcement of a stock split (Fama, Fisher, Jensen, and Roll, 1969). While much has changed over the intervening decades, the basic event study methodology used by most practitioners has changed little. In a perfectly efficient market, the price of a security reflects all available information known to the market, so in such a market the price of a security will immediately respond to the introduction of new information. After determining the firms and dates subject to an event, an analyst can determine its impact by calculating the difference between the realized return on the security, and the prediction from a model of expected returns. This difference, often called the abnormal or excess return, can be attributed to the impact of the event, conditional on the adequacy of the model generating expected returns.

Although the academic literature featuring event studies as an empirical device is long and developed, event studies by scholars writing for academic readers have been used almost exclusively to test the impact of an event on a broad cross-section of securities rather than one particular corporation’s stock (Brav and Heaton, 2015). Inference in such studies is sometimes done using flexible or nonparametric methods, but usually it is based on comparing $t$-statistics to critical values of the Student’s $t$ distribution. As (Gelbach, Helland, and Klick, 2013) point out, that approach requires one of two conditions to hold. First, if excess returns are normally distributed, the Student’s $t$ distribution is correct in finite
samples. But there is considerable evidence that excess returns are not normal. Second, if there are enough firms and dates that experience the event of interest so that a central limit theorem can reasonably be expected to usefully apply to the estimated event effect, then the estimated event effect—which is an average of a sort—will be approximately normal. But (Gelbach et al., 2013) observe that in single-firm event studies used for litigation, each date of interest is functionally an event study with only 1 firm-date combination. Consequently, the large-sample justification for standard inferential approaches also fails. The result is that the standard approach to inference yields invalid inference in single-firm/single-event studies of the sort commonly used in securities litigation.

In light of the increased use of event studies for legal and regulatory purposes, a nascent literature has developed exploring potential remedies for this and other problems. Gelbach et al. (2013) explore the “general invalidity” of standard inference in single firm event studies. Using a Monte Carlo simulation, they demonstrate that the standard approach used by most analysts performs poorly in terms of Type I and Type II error rates in the period of 2000-2007. Baker (2016) explores the empirical properties of the standard event study approach to returns on the securities of firms in the Dow 30 and S&P 500 industries during the financial crisis. He finds a consistent underestimation of standard errors in the presence of shifting market volatility and inflated test rejection rates. Brav and Heaton (2015) warn judges against having “unrealistic expectations of litigants’ ability to quantitatively decompose observed price impacts”. Finally, Fisch, Gelbach, and Klick (2018) explore the consequences of different design decisions on the Halliburton case, showing how attention to oft-ignored methodological issues can have substantial implications for case determinations.

However, this literature has dealt primarily with the inferential properties of single-firm event studies, i.e., how significance tests for event-date excess returns perform in practice.¹ This makes sense given that plaintiffs bringing securities actions under SEC Rule 10b-5 must

demonstrate reliance, materiality, and loss causation, all of which often hinge in practice on proving that price moved on dates when there were alleged material misrepresentations or disclosures of fact. As a result, the modifications to the standard approach proposed in Gelbach et al. (2013), Baker (2016) and Fisch et al. (2018) involve suggestions for more robust estimates of the variance of excess returns and/or the critical values used for testing statistical significance. However, these modifications focus little attention on the estimators of the coefficients used to calculate the event-date excess return. Given that the excess returns are the parameters that determine the damage estimates in securities suits, it is worthwhile to explore whether methods exist that can provide more accurate estimates of the excess return itself.

In this paper we reconsider the general approach to conducting event studies, focusing on whether recently (and not so recently) developed machine learning techniques can improve estimation of the expected returns. Properly viewed, an event study is in essence an out-of-sample prediction problem.

The utility of doing so may be seen by considering two possible models of expected return. Letting the measure of the daily return for firm $i$ on date $t$ be $r_{it}$ and the vector of variables used to predict $r_{it}$ be $X_{it}$, we write

$$r_{ij} = g^j(X_{it}) + \zeta_{it}^j,$$

where evaluating the function $g^j$ to $X_{it}$ yields the predicted return for model $j \in \{1, 2\}$, so that $\zeta_{it}^j$ is the model-$j$ excess return for firm $i$ on date $t$.

Now define $u_{it} = g^2(X_{it}) - g^1(X_{it})$. It follows that $\zeta_{it}^1 = \zeta_{it}^2 + u_{it}$. Suppose it is true that

---

2A different question is whether classical statistical significance testing is the right approach to assessing whether there was price impact. Work by Gelbach & Hawkins (forthcoming) addresses this question, but for now we ignore it in this paper.

3To be sure, Baker (2016) proposes an FGLS event study method that yields different coefficient estimates from the standard OLS ones. And Fisch et al. (2018) use a GARCH model, which also yields different coefficient estimates. But these differences are essentially byproducts of a focus on properly estimating second-moment properties, rather than the coefficient estimates themselves.
(i) \( u_{it} \) has positive variance and (ii) \( u_{it} \) and \( \zeta^1_{it} \) are independent. Then model 2 necessarily has lower variance, because (i) and (ii) imply \( V(\zeta^1_{it}) > V(\zeta^2_{it}) \). If \( g^2 \) and \( g^1 \) also have the same mean, then the two models have the same bias (whether or not it is zero), so model 2 also has lower mean squared error (MSE). Even if model 2 has greater bias than model 1, in which case \( E[u_{it}] \neq 0 \), it is still possible that model 2 is superior in MSE terms. Whether it is depends on whether the squared bias outweighs the reduction in variance. This paper takes seriously such possibilities by considering the MSE performance of a large variety of return models.

We note that MSE performance can be improved in two distinct ways. One is to provide a better model of the predicted return given data \( X_{it} \). That corresponds to the previous paragraph’s discussion of situations in which model 2 is better than model 1 along the MSE metric. Another way to improve MSE performance is to retain the same model, in the sense that the predicted value of \( r_{it} \) given \( X_{it} \) is the same, but to use a better way to estimate the parameters of that fixed model. A familiar example involves a linear regression model in which there is some non-sphericality, in which case one can improve on the MSE of OLS-based predictions by using a lower-variance coefficient estimator such as FGLS. We distinguish the two by using the word “specification” to refer to the combination of a model and an estimator of that model’s parameters.

Given that event study specification selection can be conceptualized as a prediction problem, there is good reason to think we can do better than the specification commonly used in securities litigation involving the OLS estimation of the simple market model. Work in computer science and statistics has consistently demonstrated that OLS overfits data when used to for prediction purposes (Tibshirani, 1996). Although OLS provides the best unbiased linear prediction in-sample, it often exchanges greater variance to achieve in-sample bias reduction, and that leads to poor prediction accuracy. Modern machine learning methods accept some bias in return for reducing variance, and they are implemented by training estimators to directly minimize out-of-sample prediction error.
Using real stock return data, we demonstrate that a number of out-of-the box statistical approaches that are relatively easy to interpret perform better than the standard, OLS-based event study models used in court proceedings. Specifications that perform well rely on penalized regression, those which optimize over a set of individual peer firm returns rather than a single pre-fab peer index, and those that involve cross-validation that is robust to otherwise unmodeled time-series properties of the data generating process. The best specifications provide substantial improvements over event study approaches often used in securities litigation. Our approach thus suggests reliability can be significantly improved with appropriate estimation techniques.

2. Prior Literature

An event study is at base a statistical procedure used to measure the effect of an unexpected event on the value of a firm’s traded security, and has been described as “one of the most frequently used analytical tools” in applied corporate finance research (Peterson, 1989). Eugene Fama, a leading figure in asset pricing and academic finance, has gone so far as to say that before the event study there was little empirical evidence on the core issues of financial economics, whereas now “we are overwhelmed with results, mostly from event studies” (Fama, 1991). Event studies are popular in large measure for their simple and elegant method of controlling for confounding market effects when modeling stock returns, which in theory allows for the isolation of causal effects of events such as regulatory changes or changes in capital structure.

Event study methodology in finance began with a paper by Fama, Lawrence Fisher, Michael Jensen, and Richard Roll in 1969. Theoretical articles by Samuelson and Mandelbrot had demonstrated that securities trading on exchanges exhibited indicia of efficiency, as reflected in their independence properties. But there had been little actual empirical evidence.

---

4We are also working on applying this insight to the many-firm studies more commonly used in academic research.
of the speed of price adjustment to specific forms of information entering the market. Fama et al. (1969) used the presence of stock splits to test whether there was “unusual behavior” in the return on a security in the months leading up to the split. Notably, the event study format they used follows the same functional form as event studies used today in court proceedings, with the log of one plus an individual security’s returns regressed on a constant and the log of one plus the return on a market index.

Following Fama et al. (1969) thousands of articles have been published in leading journals using event studies to isolate the impact of a broad range of corporate events.\(^5\) Decades later, a parallel literature developed analyzing the properties of the comparative statistical models used for event studies. A pair of articles written by Stephen Brown and Jerold Warner compared the ability of competing specifications to detect abnormal performance using both monthly and daily data (Brown and Warner, 1980, 1985). Brown and Warner’s 1985 paper, which has come to define the field, declared that event studies presented few practical difficulties when conducted using daily data. The authors acknowledged that stock returns departed from normality, but still they found OLS-based methods to be largely robust to parametric concerns. Subsequent studies tested the properties of event study methods, analyzing how frequently different tests reject the null hypothesis of no abnormal performance, and the power of specifications to detect abnormal performance when imputed (Binder, 1998).

Later empirical studies also questioned the generalizability of Brown and Warner’s results. Chandra, Moriarity, and Lee Willinger (1990) showed that the relative equivalence in performance between the OLS/market model specification and simpler approaches was a statistical artifact of specification implementation. Moreover, subsequent research verified that excess returns were not normally distributed, and suggested that for outlier-prone data (as is prevalent in financial markets), the true Type I error rate will be larger than that associated with asymptotic values, particularly for stocks with high kurtosis (Hein and West-
fall, 2004). Some scholars proposed using non-parametric tests of abnormal performance to address non-normality in many-firm studies, e.g., rank and sign tests (Corrado, 1989).

Recently, scholars have scrutinized the application of academic event studies in litigation. Corrado (2011) notes that single-security event studies rarely arise in academic literature but are routinely proffered as evidence in court proceedings. He advises legal practitioners to use a simple nonparametric modifications to the event study procedure that would at least correct for the non-normality of individual stock returns. Gelbach et al. (2013) propose a similar modification they termed the SQ test. To perform a lower-tailed version of this test with classical significance level $\alpha$, one ranks the estimated excess returns from the market model regression and determines whether the event-date excess return is more negative than the $\alpha$-quantile of the empirical distribution of estimated excess returns from the pre-event window.\footnote{For an upper-tailed version, one determines whether the event-date excess return is greater than the $(1 - \alpha)$-quantile; for a two-sided version, one determines whether the event-date excess return is between the $(\alpha/2)$-quantile and $(1 - \alpha/2)$-quantile.} Using a dataset containing the returns for all securities in the Center for Research in Security Performance’s (CRSP) database from 2000 to 2007, the authors uncover substantial evidence of bias against finding statistically significant excess returns.

Baker (2016) analyzes the performance of a group of event study specifications over the financial crisis period of 2007-2009. He finds that when volatility in the market shifts suddenly, standard specifications with a constant estimation period and variance estimate will fail to reflect the changed nature of stock returns. As a proposed remedy he suggests using either feasible generalized least squares (FGLS) or an estimator that adjusts the standard error of the t-statistic by the ratio of changes in market volatility to more fully reflect the true variance of the residuals from the market model. Fisch et al. (2018) propose dealing with this same issue using a generalized autoregressive conditional heteroskedasticity (GARCH) estimator for the variance of daily returns and then using daily estimates of the variance to obtain a normalized white noise term to which the SQ test may be applied.

However, it is important to note that none of the proposed remedies described above
fundamentally changes the estimation approach taken to predict the event-date excess return itself. This is the province of the present paper.

3. Methodology

The steps necessary to conduct an event study have not changed substantially since Fama et al. (1969). An analyst must first identify a return series covering the event at issue, ensure that the stock trades frequently enough for each return to cover only one day (or at most a few days), and establish the dates on which the event occurred. There are then three steps to conducting an event study: (1) defining the “event window,” (2) calculating the excess return of the stock over the event window, and (3) testing for statistical significance of the excess return.

The event window is the period over which the impact of the event will be tested. Because event studies are built upon the underpinnings of the efficient markets hypothesis, the typical presumption is that stock price will quickly adapt to new information released to the market. As a result, event windows (especially those used in litigation) are typically short, perhaps as short as the one-day trading period surrounding an event. Occasionally the event window may be extended to multiple days, particularly if the time that the information was released to the market is uncertain, or if there is reason to believe that the information was unlikely to be quickly absorbed into the stock price (Mitchell and Netter, 1994). Extending the event-window length does risk reducing power, and it may compromise the event study’s ability to identify abnormal performance.

After defining the event window, it is necessary to isolate the portion of the security return attributable to the new information from general fluctuations in stock price. This is the primary role of the event study: to determine whether estimated effects fall outside the range that would be expected due to the usual variation in the stock’s returns. The most

---

7For an interesting example of a case in which the speed of capitalization was at issue, see In re Apollo, Inc., Securities Litigation. ADD MORE.
significant determination is in the method selected to determine the expected return. Model variants are generally divided into two categories: “statistical” and “economic”. Statistical models rely only on the empirical properties of asset returns, while economic models bring in additional assumptions on investor behavior (Mackinlay, 1997). While the original development of the event study technique was built upon the theoretical foundation of the Capital Asset Pricing Model (CAPM), most finance scholars no longer consider the CAPM to be an accurate model of price formation (Fama and French, 1996). Modern additions to the standard market model event study, including the factor based approaches popularized by Fama-French and Carhart, are based upon the predictive power of security features rather than any theoretical justification. It has been argued that economic models impose additional statistical assumptions without offering many practical advantages (Campbell, Lo, and Mackinlay, 1997).

We can parsimoniously write a given model of expected returns as:

\[ r_{it} = g(X_{it}) + \zeta_{it}, \]  

(1)

where \( r_{it} \) is the measure of the daily stock return, \( g \) is some function that captures the details of the model in question, and \( \zeta_{it} \) is the excess return on date \( t \) for firm \( i \). The return on the stock \( r_{it} \) is equal to the expected return \( g(X_{it}) \) plus a noise component \( \zeta_{it} \). If we assume that the noise is mean 0, then the expected return is

\[ E[r_{it}] = g(X_{it}) \]

For the market model, the expectation function \( g \) is assumed linear in the market return:

\[ g(\cdot) = \alpha + \beta M_t \]

Because the excess return for a stock on date \( t \) is simply the difference between the realized
return $r_{it}$ and the expected return, the excess return is

$$\zeta_{it}^{MM} = r_{it} - g^{MM}(X_{it})$$  \hspace{1cm} (2)$$

This equation can be viewed as a prediction problem for which modern machine learning methods have proven quite adept (Kleinberg, Ludwig, Mullainathan, and Obermeyer, 2015). When we view an event study as a prediction problem, our goal is to isolate the portion of the return that cannot be predicted by other explanations. This amounts to finding the $g(X_{it})$ with the highest predictive power according to a chosen metric (we generally use MSE as that metric). There is little reason a priori to assume that the OLS/linear market model specification is the one that minimizes this prediction error.

Below we consider 31 specifications. Of these, 15 include the Fama-French and Carhart factors, and 16 do not. Indexing each specification with $k$, we may write the $k^{th}$ specification as $r_{it} = g^k(X_{it}) + \zeta_{it}^k$. For each of 10,000 simulation replicates, we pick a random firm $i$ and a random date $t$, which we think of as the event date. We also pick 250 dates for firm $i$, which are used as the estimation set for each return specification. How we pick the 250 dates varies a bit; see the discussion below. For each replication $b \in \{1, 2, \ldots, 10,000\}$ of the simulation study we describe below, we thus obtain 31 estimates of the excess return for the event date. Write $it(b)$ to indicate the firm-date used as the event-date for the $b^{th}$ simulation replicate.

For each $it(b)$, we obtain 31 estimates $\hat{\zeta}_{it(b)}^k$, $k \in \{1, 2, \ldots, 31\}$. Let $w_b^k \equiv (\hat{\zeta}_{it(b)}^k)^2$ be the squared value of the estimate for the $b^{th}$ replication of the $k^{th}$ specification. For each simulation replicate $b$, we compute the mean and standard deviation across these 31 estimates of $w_b^k$; call these $\bar{W}_b$ and $SD_b$. Then we standardize the 31 values of $w_b^k$ by subtracting $\bar{W}_b$ from each and dividing the result by $SD_b$: $Z_b^k \equiv (w_b^k - \bar{W}_b)/SD_b$.

Finally, for each specification $k$, we compute the mean value of $Z_b^k$ over all 10,000 replicates. The result is $\bar{Z}^k$, which is what we plot in the graphs reporting our results. We can interpret each $\bar{Z}^k$ estimate as a relative measure of the magnitude of prediction error ob-
tained using each of the 31 specifications. A specification with $Z^k = 0$ performs at an average level. The specification with the highest $Z^k$ is the worst (most error), and the specification with the lowest $Z^k$ is the best (least error).

3.1. Specifications

In this study we compare the predictive power of alternative expectation functions using real stock returns from CRSP over the period from 2009 to 2019. Below we describe the various specifications used for this comparison, which are all straightforward extensions of models and estimation methods used in common practice. For all specifications we use log-transformed return series, where $r_{it} = \ln P_t - \ln P_{t-1}$. The log transform is desirable because it tends to remove at least some non-normality in stock returns.

Define $\zeta_{it}(b) = r_{it} - X_{it}'b$. This is the excess return function for any linear model of returns using data $X_{it}$, given that the coefficient vector used is $b$. Also define $\zeta_i(b)$ as the column vector of $\zeta_{it}(b)$ values over the dates used for firm $i$ (we use 250 non-event dates plus one event date, so $\zeta_i(b)$ has length 251). Most specifications we consider are based on minimizing a function that is or includes the sum of squared errors, which means the errors enter quadratically via $\zeta_i(b)'\zeta_i(b)$. Other specifications are based on minimizing absolute errors; see below for more.

3.1.1. Specification 1 - Market Model (MM)

This is the basic market model approach used widely in academic research and by experts in litigation. It models the return on a stock as a function of the return on a market index. Here we use the return on the S&P 500 Index as a proxy for aggregate movement in the stock market. The model for the 250-day estimation window is:

$$r_{it} = \alpha^{MM} + \beta^{MM} mkt_{it} + \zeta_{it}^{MM},$$

(3)
with $E[\zeta_{it}^{MM} | mkt_{it}] = 0$. For date 251, the excess return is $r_{i,251} = [\hat{\alpha}^{MM} + \hat{\beta}^{MM} \times mkt_{i,251}]$, where $(\hat{\alpha}^{MM}, \hat{\beta}^{MM})$ is the vector of OLS estimates of the coefficients in equation (3).

3.1.2. Specification 2 - Market Model + Peer Index (MMPI)

In this simple extension to Specification 1, we add as a regressor the equally-weighted return index from firms in the same SIC industry as firm $i$ (such a peer index is commonly used in litigation). The peer index is constructed from all firms in the same 4-digit SIC industry as firm $i$, unless there are fewer than eight such firms, in which case it is constructed using all firms in the same 3-digit SIC industry as $i$.\(^8\) The return model is is:

$$r_{it} = \alpha^{MMPI} + \beta_1^{MMPI} mkt_{it} + \beta_2^{MMPI} peer_{it} + \zeta_{it}^{MMPI}, \quad (4)$$

and the excess returns are $r_{i,251} = [\hat{\alpha}^{MMPI} + \hat{\beta}_1^{MMPI} \times mkt_{i,251} + \hat{\beta}_2^{MMPI} \times peer_{i,251}]$, where $(\hat{\alpha}^{MMPI}, \hat{\beta}_1^{MMPI}, \hat{\beta}_2^{MMPI})$ is the vector of OLS estimates of the coefficients in equation (4).

3.1.3. Specification 3 - Median Regression (Med)

Koenker and Bassett (1978) show that estimation based on objective functions based on MSE can be importantly sensitive to modest amounts of outlier “contamination”. Where residuals are non-Gaussian and long-tailed, MSE-based estimators can be “very poor”. It is now a well-established fact that stock returns don’t derive from the normal distribution and have excess mass in the tails of the distribution (e.g. Gelbach et al. (2013) and references therein). Koenker and Bassett (1978) suggest various quantile regression-based estimators in such situations.

Specification 3 thus uses median regression to estimate the coefficients in specification 2. Median regression can be understood equivalently as (i) choosing $b$ to minimize $\sum_{t=1}^{250} |\zeta_{it}^{MMPI}(b)|$, or (ii) choosing $b$ to minimize $\sum_{t=1}^{250} \rho_{0.5}(\zeta_{it}^{MMPI}(b))$, where $\rho_r$ is the check function defined so that $\rho_r(\zeta) = \zeta[\tau - 1(\zeta < 0)]$. The excess returns for specification 3 are

\(^8\)If there are fewer than five such firms we drop them from consideration.
\[ r_{i,251} - [\hat{\alpha}^{Med} + \hat{\beta}_1^{Med} \times mkt_{i,251} + \hat{\beta}_2^{Med} \times peer_{i,251}], \] where \((\hat{\alpha}^{Med}, \hat{\beta}_1^{Med}, \hat{\beta}_2^{Med})\) is the vector of estimated median regression coefficients.

3.1.4. Specification 4 - Gastwirth Regression (GR)

This specification is based on the proposal by Gastwirth\(^9\) to estimate regression coefficients using estimated median regression coefficients together with estimates based on the 1/3- and 2/3-quantiles. To implement it, let \(\hat{\beta}(\tau)\) be the vector of estimated \(\tau\)-quantile regression coefficients for each choice of \(\tau \in \{1/3, 1/2, 2/3\}\). Following Gastwirth, we then calculate the weighted average

\[ \hat{\beta}^{GR} \equiv 0.1(3\hat{\beta}(1/3) + 4\hat{\beta}(1/2) + 3\hat{\beta}(2/3)) \]

Estimated excess returns in the Gastwirth specification are then calculated as \(r_{it} - X'_it \hat{\beta}^{GR}\).

3.1.5. Specification 5 - Trimmean Regression (Trimmean)

This is the final quantile regression model proposed in Koenker and Bassett (1978). It is similar to the Gastwirth specification, Specification 4. Instead of using the 1/3, median, and 2/3 quantile regression estimates, though, Specification 5 uses the “trimmean” estimator based on the 1/4, median, and 3/4 quantiles. The estimated coefficient vector is

\[ \hat{\beta}^{Trimmean} \equiv .25[\hat{\beta}(1/4) + 2\hat{\beta}(1/2) + \hat{\beta}(3/4)]. \]

Estimated excess returns in the Gastwirth specification are then calculated as \(r_{it} - X'_it \hat{\beta}^{Trimmean}\).

3.1.6. **Specification 6 - Elastic Net Regularization with 2 Factor Model (ENR)**

We return now to MSE-based objective functions, introducing our first regularized regression estimator. Regularized regression alters the least-squares objective function by imposing a penalty on coefficient magnitude. This has the effect of reducing overfitting with in-sample data.

We use a form of penalized regression objective function known as elastic net regularization. This form allows weight on both the sum of squared coefficients and the sum of their absolute value. Assuming there are $p$ coefficients to estimate, elastic net regularization entails choosing coefficients $c$ to minimize the objective function

$$
\zeta_i(c)'\zeta_i(c) + \frac{1}{2}\lambda \left( [1 - a] c'c + 2a \sum_{j=1}^{p} |c_j| \right),
$$

where $a$ and $\lambda$ are regularization parameters to be chosen as part of the estimation. When $a = 1$, elastic net regularization is equivalent to lasso regression, which tends to set many coefficient estimates to zero (for this reason lasso is often used for model selection). When $a = 0$, elastic net regularization is equivalent to ridge regression, which tends to push coefficient estimates toward each other.

We don’t have strong priors on whether lasso or ridge penalties are more appropriate, so instead of choosing a value of $a$ a priori, we optimize over it in the estimation. To obtain our elastic net regularization estimates of model mean squared error, we do the following for each set of 251 observations:

- Randomly group the data into ten random groupings known as “folds”.
- For each of the nine values of $a \in \{0.1, 0.2, \ldots, 1\}$ use cross-validation to find the minimizing value of $(c', \lambda, a)$; call the estimate of $\lambda$, $\lambda^*(a)$.
- Then select the value of $a$ that yields the lowest MSE among the nine MSE-minimizing values; call this value of $a$, $a^*$. 
• Set $\hat{\beta}^{ENR}$ equal to the estimates of the $c'$ coefficients from the model with $a = a^*$ and $\lambda = \lambda(a^*)$.

• Calculate excess returns as $\zeta_{it}(\hat{\beta}^{ENR})$.

3.1.7. Specification 7 - Elastic Net Regularization with Unconstrained Peer Firm Returns (ENR-U)

This specification generalizes specification 6 by relaxing the constraint that peer firms' returns enter the model through an equally weighted returns index. Specification 7 drops that constraint and estimates a distinct coefficient for each peer firm's daily return. Notice that this specification nests specification 6, because we obtain the peer firm index by setting the coefficient on each of the $N_{\text{peer}}$ peer firms' returns equal to $N_{\text{peer}}^{-1}$. Thus this specification involves a strictly more flexible model than all the other models whose regressor set is the market return and peer index.\textsuperscript{10}

For specification 7, we implement the unconstrained peer firm returns model using the same elastic net regularization approach as in specification 6. This means the vector of coefficients $c$ used to calculate estimates of $\zeta_{it}(c)$ has $2 + N_{\text{peer}}$ dimensions—one for $a$, one for $c_1$, and one for each of the $N_{\text{peer}}$ firm returns.

3.1.8. Specification 8 - Regularization All Peer Firms and Forced Market Inclusion (ENR-FMI)

This specification augments specification 7 by forcing the estimation process to include the broad market index variable. That is, specification 8 differs from specification 7 because specification 7 can choose to drop the market index regressor before calculating coefficient estimates, whereas specification 8 must include it in the regressor set. We investigate this

\textsuperscript{10}Note that if a firm were to have more than 250 peer firms, including each firm individually would be impossible. This is another way in which penalized regression is useful. Penalization allows for more covariates than observations because it drops weakly correlated controls from the estimation equation. This helps explain why lasso is frequently used for model selection problems.
specification out of a belief that some experts and courts might insist that the market index be part of the model used to predict excess returns. Calculating the ENR-FMI coefficient estimate is done using the same method as in specification 8, but with the penalty terms being $c'c - (c_1^{ENR-FMI})^2$ and $(\sum_p |c_p|) - c_1^{ENR-FMI}$.

3.1.9. Specification 9 - Two-Factor Model with Lasso-Based Equally Weighted Index (2FM-LEW)

This specification is like specification 2 in that its last step involves using OLS to estimate a two-factor model that includes the market index and an equally weighted peer index. It differs from specification 2 in that the peer firms included in that index are selected using lasso. The first step of specification 9 can be thought of as a version of specification 6’s elastic net regularization, except with $a$ set to 1 so that this step involves lasso rather than a blend of lasso and ridge regression. Once we have a set of selected peer firms—that is, once we see which peer firms the lasso model has given nonzero coefficients—we use them to calculate the equally weighted peer index. Then we do the last step mentioned above (OLS estimation of the two-factor model with the lasso-selected peer firms). In any simulation replication on which no peer stocks are given non-zero weights, the last step uses OLS estimation of the one-factor market model.

3.1.10. Specification 10 - Two-Factor Model with Median Regularization (2FM-Med)

This specification blends specifications 3 and 6. The regressor set includes the market index and the peer index, as in both those specifications. We use elastic net regularization in this specification, as in specification 6. But instead of using the quadratic term $\zeta_{it}(c)'\zeta_{it}(c)$ in the objective function as in specification 6, we use the sum of absolute errors, $\sum_{t=1}^{250} |\zeta_{it}(c)|$, as in specification 3’s median regression approach.
3.1.11. Specification 11 - Median Regularization with Market Index and All Peer Firms
(ENR-U-Med)

This specification starts from specification 7, regularized regression with the regressor set including the market index and all peer firms entering individually. The difference is that in specification 11 we use the sum of absolute errors, \( \sum_{t=1}^{250} |\zeta_{it}(c)| \), rather than the sum of squared errors.\(^\text{11}\)


This specification starts from specification 8, regularized regression with (i) forced market index inclusion and (ii) all peer firms entering individually. The difference is that in specification 12 we use the sum of absolute errors, \( \sum_{t=1}^{250} |\zeta_{it}(c)| \), rather than the sum of squared errors.\(^\text{12}\)

3.1.13. Specification 13 - Two-Factor Local Linear Random Forest Regressions (LLRF-2F)

Random forests, as described in Breiman (2001), are a popular method for non-parametric regression that have proven effective across a range of applications. Random forests are popular, in part, because they require little model tuning and perform well “out-of-the-box” compared to more complex machine learning methods like neural nets. Random forests are forms of regression trees that are notably effective when used with a large number of features that are unelated to the outcome variable, as is likely the case in event studies. Here we use the local linear random forest method from Friedberg, Tibshirani, Athey, and Wager (2018). This method takes the perspective of random forests as an adaptive kernel method, and pairs the forest kernel with a local linear regression adjustment to better capture smoothness. Specification 13 uses local linear random forests with two factors: a market and

\(^{11}\)Thus this specification alters specification 7 in the same way that specification 10 alters specification 6.

\(^{12}\)Thus this specification alters specification 8 in the same way that specification 10 alters specification 6 and specification 11 alters specification 7.
peer index.

Returns (LLRF-U)**

This model starts from specification 13 and then relaxes it by allowing each peer return to enter individually, rather than forcing them to enter via an equally-weighted index.

3.1.15. **Specification 15 - Two-Factor Time-Series Cross-Validation (TSCV-2F)**

The penalized regressions described above estimate the penalization parameters $\alpha$ and $\lambda$ through conventional cross-validation. That method is not always optimal with time series data, as it ignores any trend component to the relationships. Various alternative cross-validation techniques have been proposed to address this issue. Specification 15 uses the “evaluation on a rolling forecasting origin” method.

In this procedure, a series of test sets consisting of a single observation is used for cross-validation. The corresponding training set consists of only those observations that occurred prior to the observation that forms the test set (with a floor of at least 50 observations). Thus, no future observations are used in constructing the forecast. The following diagram illustrates the series of training and test sets. Blue observations (to the left, for those reading in black and white) form the training sets; each red observation that immediately follows a set of training observations forms a test set (the gray observations to the right of each red one are left out). Forecast accuracy is computed by averaging over test sets.

Specification 15 includes two factors and the elastic net regularization penalty function but with time-series cross-validation. Thus it is the same objective function as in specification 6, but with time-series cross-validation.
3.1.16. *Specification 16 - Time-Series Cross-Validation with Market Index and All Peer (TSCV-U)*

This specification is the same model as specification 15, except that each peer firm's return is allowed to enter individually rather than as an equally-weighted index. Thus it can also be viewed as specification 7 but with time series cross-validation.

For referential convenience, Table 3.1.16 provides a list of specification acronyms, numbers, and descriptions.
<table>
<thead>
<tr>
<th>Model Acronym</th>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>1</td>
<td>Market model</td>
</tr>
<tr>
<td>MMPI</td>
<td>2</td>
<td>Market model with peer index</td>
</tr>
<tr>
<td>Med</td>
<td>3</td>
<td>Median regression</td>
</tr>
<tr>
<td>GR</td>
<td>4</td>
<td>Gastwirth regression</td>
</tr>
<tr>
<td>Trimmean</td>
<td>5</td>
<td>Trimmean regression</td>
</tr>
<tr>
<td>ENR</td>
<td>6</td>
<td>Elastic Net Regularization with 2 Factor Model</td>
</tr>
<tr>
<td>ENR-U</td>
<td>7</td>
<td>Elastic Net Regularization with Unconstrained Peer Firm Returns</td>
</tr>
<tr>
<td>ENR-FMI</td>
<td>8</td>
<td>Regularization All Peer Firms and Forced Market Inclusion</td>
</tr>
<tr>
<td>2FM-LEW</td>
<td>9</td>
<td>Two-Factor Model with Lasso-Based Equally Weighted Index</td>
</tr>
<tr>
<td>2FM-Med</td>
<td>10</td>
<td>Two-Factor Model with Median Regularization</td>
</tr>
<tr>
<td>ENR-U-Med</td>
<td>11</td>
<td>Median Regularization with Market Index and All Peer Firms</td>
</tr>
<tr>
<td>ENR-FMI-U-Med</td>
<td>12</td>
<td>Median Regularization with Forced Market Index Inclusion and All Peer Firms</td>
</tr>
<tr>
<td>LLRF-2F</td>
<td>13</td>
<td>Two-Factor Local Linear Random Forest Regressions</td>
</tr>
<tr>
<td>LLRF-U</td>
<td>14</td>
<td>Local Linear Random Forest Regressions With Market and Peer Returns</td>
</tr>
<tr>
<td>TSCV-2F</td>
<td>15</td>
<td>Two-Factor Time-Series Cross-Validation</td>
</tr>
<tr>
<td>TSCV-U</td>
<td>16</td>
<td>Time-Series Cross-Validation with Market Index and All Peer</td>
</tr>
</tbody>
</table>
4. Simulation Procedure and Results

To test the relative predictive accuracy of the sixteen specifications described above, 10,000 unique firm-events are selected at random over the period from 2009 to 2019 in the CRSP dataset. As is common in the literature, we exclude all unit investment trusts (SIC 6726), real estate investment trusts (SIC 6798), and non-identifiable establishments (SIC 9999). When selecting random event dates, the security in question is required to have a complete return series for the 250 trading dates directly preceding the event in question. As mentioned above, in selecting peers, only other firms with complete return series over the same period in the same four-digit SIC industry are used. If there are fewer than eight such firms, we use peers in the same three-digit SIC industry.

We now explain how we calculated our results for the 31 specifications, with and without the Fama-French/Carhart factors. Indexing each specification with \( k \), we may write the \( k^{th} \) specification as 
\[
r_{it} = g^k(X_{it}) + \zeta^k_{it}.
\]
For each of 10,000 simulation replicates, we pick a random firm \( i \) and a random date \( t \), which we think of as the event date. We also pick 250 dates for firm \( i \), which are used as the estimation set for each return specification. How we pick the 250 dates varies a bit; see the discussion below. For each replication \( b \in \{1, 2, \ldots, 10,000\} \) of the simulation study we describe below, we thus obtain 31 estimates of the excess return for the event date. Write \( it(b) \) to indicate the firm-date used as the event-date for the \( b^{th} \) simulation replicate.

For each \( it(b) \), we obtain 31 estimates \( \hat{\zeta}^k_{it(b)} \), \( k \in \{1, 2, \ldots, 31\} \). Let 
\[
w^k_b \equiv (\hat{\zeta}^k_{it(b)})^2
\]
be the squared value of the estimate for the \( b^{th} \) replication of the \( k^{th} \) specification. For each simulation replicate \( b \), we compute the mean and standard deviation across these 31 estimates of \( w^k_b \); call these \( \bar{W}_b \) and \( SD_b \). Then we standardize the 31 values of \( w^k_b \) by subtracting \( \bar{W}_b \) from each and dividing the result by \( SD_b \): 
\[
Z^k_b \equiv (w^k_b - \bar{W}_b)/SD_b.
\]

Finally, for each specification \( k \), we compute the mean value of \( Z^k_b \) over all 10,000 replicates. The result is \( \bar{Z}^k \), which is what we plot in the graphs reporting our results. We can interpret each \( \bar{Z}^k \) estimate as a relative measure of the magnitude of prediction error.
obtained using each of the 31 specifications. A specification with $Z^k = 0$ performs at an average level. The specification with the highest $Z^k$ is the worst one (most error), and the specification with the lowest $Z^k$ is the best one (least error).

This standardization approach is useful because the randomly selected “event” dates vary across firm and date, so that some “event” dates come from series with considerably greater variation than others. Standardizing in the way we do puts each date on par with the others (at least up to second-moment properties). In Figure 1 below, we plot the mean and standard deviation of the normalized prediction errors of the 16 models without the Fama-French Carhart factors.

Figure 1. Mean Squared Error By Specification: 10,000 Simulations

Note: Figure 1 plots the average normalized squared prediction error for our 16 candidate specifications, with squared prediction errors normalized date-by-date. An average of zero indicates that a specification has average predictive power, and better performance is associated with lower values in the plot.
As evidenced in Figure 1, an immediate implication from these simulation results is that the one-factor market model performs quite poorly when compared to the other specifications. While studies have suggested that a one-factor specification is often sufficient to predict returns, (Baker, 2016), the results reported here indicate cause for concern. In addition, we find support for the proposition in Koenker and Bassett (1978) that quantile regression is a more robust method for estimating conditional expectations with fat-tailed data: all three quantile-based specifications (specifications 3-5) outperform the non-regularized OLS-based methods.

The penalized regressions also outperform the OLS specifications, particularly when the penalization specification is estimated over the individual peer returns rather than the equally weighted index. This suggests that allowing the data-adaptive attributes of regularized specifications to determine the important peer firms is a potentially more promising approach than having each litigant’s analyst select a peer index based on subjective determinations of comparability. The best performing specification in our sample was time-series cross-validation to estimate the penalization parameters. This indicates that the time-series attributes of stock return data may be important than is often thought.

Finally, we note that specifications 9-13 perform worse than the simple OLS/two-factor specification approach that includes the equally weighted peer index. This raises doubts about the performance of penalized regression when used with quantile regression, although perhaps it is unsurprising that a quantile-based objective function does worse when measured against squared-error performance. Perhaps quantile-based specifications would outperform the squared-error specifications if we measured in terms of absolute rather than squared errors.

Another way to explore the results is to examine the entire distribution of the normalized deviations, rather than only the arithmetic means. Figure 2 plots the kernel density estimates of the distribution of normalized prediction errors by specification. The densities reveal a similar story to the means: the one-factor model performs generally poorly, having a
flatter distribution than the other models. However, the poor performance of the quantile penalization specifications (10-12) is driven by a bimodal pattern in the relative densities, with greater mass on both positive and negative deviations when compared to the simple two-factor specification. Quantile penalization specifications also produce higher variance estimates.

Figure 2. Standardized Density by Specification: 10,000 Simulations

Note: Figure 2 plots the normalized densities of the prediction error for our sixteen candidate specifications using a kernel density estimate. The squared prediction errors are normalized by date, so that each value is the relative comparison across specifications. A mass of weight at zero would suggest that has average predictive power.

4.1. Fama-French Carhart Factors

Another question often proposed in the literature is the extent to which the “Fama-French” factors must be controlled for when running individual security event studies. Ac-
cording to the CAPM, the only significant factor in explaining the cross-section of returns is the sensitivity of a firm’s equity price to the contemporaneous return on the market. However, as demonstrated in Fama and French (1996), there is persistent evidence that other risk factors explain returns, and that the slope of the regression of a security’s return on the market index ($\beta$) does not suffice to explain expected returns. A series of papers by Fama and Ken French supported including two additional variables, involving the returns on long-short portfolios of securities sorted along size and valuation metrics. In addition, the momentum factor proposed by Carhart (1997) is often included. This momentum factor is based on the notion that there is short-term serial correlation in the market, where stocks that have recently over-performed the market will continue to overperform the market. This factor is similarly measured through a long-short portfolio of firms sorted by recent stock market performance. Although it is rarely used in single-firm event studies for litigation purposes, the Fama-French/Carhart “four-factor” model has been a workhorse of academic finance.

In Figure 3 we plot similar estimates for augmented versions of our earlier specifications, but with the three Fama-French/Carhart factors added to the set of explanatory variables. Because the one-factor specification—i.e., MM—produces substantially inferior results (it has the greatest MSE in Figure 1), we exclude it from Figure 3. We do, however, include in Figure 3 the normalized prediction error from specification MMPI, specification (MMPI) is which is the two-factor OLS event study specification without the FF/Carhart factors. The slight difference in performance between specifications (MMPI) and (MMPI_{FFC}) indicates that the Fama-French/Carhart factors reduce prediction error in excess returns, at least relative to the OLS specification.

There are several take-home points from Figure 3. First, the unrestricted time-series cross-validation specification, TSCV-U, does best among the Figure 3 specifications that include the Fama-French/Carhart factors—just as it did among those in Figure 1. Second, it
appears that the Fama-French/Carhart factors have explanatory power for some methods.\textsuperscript{13} The two best performing methods in both Figure 1 and Figure 3 were TSCV-U and ENR-U (unrestricted elastic-net regression), and the mean-squared prediction error is further from the average for both when the Fama-French/Carhart factors are included than when they are not. Some of the worst-performing specifications appear to improve, perhaps substantially, when the Fama-French/Carhart factors are included.

Third, not all methods see a reduction in mean-squared prediction error when the Fama-French/Carhart factors are included. For example, elastic-net regression with forced market-index inclusion (ENR-FMI) has greater mean-squared prediction error in Figure 3 than in Figure 1.

Finally, we observe that the variation across specifications in mean-squared prediction error performance roughly persists when the Fama-French/Carhart factors are included. One way to see this is to observe that the range of mean-squared prediction error in Figure 3 values is roughly equal to the range for the best 14 methods in Figure 1. Thus, although performance with the Fama-French/Carhart factors significantly improves by comparison to the two worst Figure 1 specifications, including the Fama-French/Carhart factors hardly eliminates the variation in mean-squared prediction error across estimation methods.

5. Conclusion

In this paper we explore whether machine learning techniques popular in other disciplines can be used to enhance the predictive power of excess return calculations in event studies conducted on single securities for securities litigation. Event studies have been used extensively in research, and the academic consensus is that they are powerful tools for detecting the impact of events on the price of a firm’s securities. However, the large sample properties that make event studies a reliable statistical tool are non-applicable when applied to the returns of a single security. In spite of these inherent weaknesses in inference, single-firm event

\textsuperscript{13}We note, though, that we have not yet done any formal hypothesis testing about this issue.
Figure 3. Standardized Density by Specification: 10,000 Simulations

Note: Figure 3 plots the average normalized squared prediction error for our 16 candidate specifications, the last 15 of which correspond to specifications 2-16 from Figure 1 augmented to include the Fama-French/Carhart factors. An average of zero indicates that a specification has average predictive power, and better performance is associated with lower values in the plot.

studies are widely used in civil litigation, with billions of dollars in settlements ultimately hinging on the outcome of a potentially flawed exercise.

Prior work has explored modifying the standard single-firm event study design to provide more robust statistical inference, but to date little work has been done to determine whether other specification variants can directly increase the precision of the excess return estimate. Given that the excess return estimate is ultimately used to calculate damages, bias in that estimate will induce biased damage figures. We take a prediction approach to the excess return calculation and find substantial performance improvement from modern
machine learning methods.
References


Fama, E. F., French, K. R., 1996. The CAPM is Wanted, Dead or Alive. THE JOURNAL OF FINANCE LI.


