Machine Learning and Predicted Returns for Event Studies in Securities Litigation

—Preliminary and Incomplete—

Andrew Baker
Stanford University

Jonah B. Gelbach
University of California at Berkeley

September 3, 2019

Abstract

We investigate the use of machine learning (ML) and other robust-estimation techniques in 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. We find that regularization (equivalently, penalized estimation) can yield noticeable improvements in both the variance of event-date excess returns and significance-test power. Thus we think there is a role for ML methods in event studies used in securities litigation. At the same time, we find that ML-induced performance improvements are smaller than those based on other good practices. Most important are (i) the use of a peer index based on returns for firms in similar industries (how this is computed appears to be less important than that some version be included), and (ii) for significance testing, using the SQ test proposed in Gelbach, Helland, and Klick (2013), because it is robust to the considerable non-normality present in excess returns.

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 are meant 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 overwhelmingly to test the impact of events on a broad cross-section of securities, rather than on 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 et al. (2013) point out, that standard approach is justified only if at least one of two conditions holds. 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) use Monte Carlo simulation to 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 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

[1] An exception is Dove, Heath, and Heaton (2019), who focus on issues involved in damages estimation.

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.[2] However, these modifications focus little attention on the estimators of the coefficients used to calculate the event-date excess return.[3] 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.[4]

Event studies can be viewed as out-of-sample prediction problems. This is important because modern machine learning (ML) methods have proven quite useful for such problems; see, e.g., Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015). In this paper, we consider whether recently (and not so recently) developed machine learning techniques can improve estimation of expected returns in relevant metrics.

To illustrate the utility of doing so, consider two possible candidate specifications for estimating the expected return, indexed by $j \in \{1, 2\}$. Let the measure of the daily return for firm i on date t be r_{it}, and let the vector of variables used to predict r_{it} be X_{it}. These predictor variables typically include the market return and might also include the four Fama-French and Carhart factors, as well as any other variables that might be used by a sufficiently

[2] A 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.

[3] To 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.

[4] Again see Dove et al. (2019)

flexible prediction function.[5] The key notational point is that every specification of the prediction function, g^j, can be viewed as mapping from the full set of predictors to the daily return value, even if some g^j functions effectively ignore some predictors.

With ζ_{it}^j defined as the excess return–equivalently, the prediction error–based on specification j, we have

$$r_{it} = g^j(X_{it}) + \zeta_{it}^j. \tag{1}$$

Now define the difference in predicted returns for the two specifications as $u_{it}(X_{it}) \equiv g^2(X_{it}) - g^1(X_{it})$; when there is no cause for confusion we suppress the argument of u_{it}. Plugging the definition of u_{it} into (1) yields the identity that relates the two excess return estimators for firm i on date t: $\zeta_{it}^1 = \zeta_{it}^2 - u_{it}$. Suppose it is true that (i) u_{it} has positive variance and (ii) u_{it} and ζ_{it}^2 are independent. Then using variance as the metric, specification 1 is noisier than specification 2: $V(\zeta_{it}^1) > V(\zeta_{it}^2)$. If g^2 and g^1 also have the same conditional mean, then the two specifications have the same bias (whether or not it is zero), so specification 2 also has lower mean squared error (MSE). But even if specification 2 has greater absolute bias than specification 1, in which case $|E[u_{it}]| > 0$, as long as this bias difference isn't too great, specification 2 will be superior in MSE terms. This is the logic of using MSE as the basis for measuring prediction accuracy, and it is the reason why ML estimators might outperform conventional least squares estimators along that metric. This paper takes

[5] According to the CAPM model, 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 (β) 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, and we conduct all our simulations both with and without these factors.

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 functional form of the predicted return given data X_{it}. That corresponds to the previous paragraph's discussion of situations in which specification 2 is better than specification 1 along the MSE metric. Another way to improve MSE performance is to retain the same functional form, so that the same underlying *model* is assumed, 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. A second example–one that we consider in this paper–is the use of quantile regression-based estimation when there is non-normality in the residuals (here, excess returns). To avoid conflating the distinct concepts of model, parameters of a functional form that is to be estimated, and estimation of those parameters, we use the word "specification" to refer to the combination of all three.

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 does so at the price of greater variance out of sample, which can lead to comparatively poor prediction accuracy in the MSE metric. Modern machine learning methods accept some bias in return for reducing out-of-sample variance. They do this 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 specifications used in court proceedings.[6], [7] We find that specifications using penalized regression generally perform well.[8] Specifications that adjust for daily market performance using data-driven peer indexes also generally perform well. Finally, we obtain generally good performance from specifications that use a cross-validation technique that's robust to otherwise unmodeled time-series properties of the data generating process. The best specifications provide noticeable improvements over event study approaches conventionally used in securities litigation.

Although we have not so far conducted any formal tests, a summary measure is the relative out-of-sample MSE of predicted excess returns for the best-performing model to the simplest "market model" specification (which happens to be the generally worst-performing specification). The best-performing specification makes use of both penalized regression and data-driven peer firm choice. Its out-of-sample variance of predicted excess returns is about 87-88% of the out-of-sample variance of predicted excess returns for the simplest market model. Given the significance recently attached to variance by Dove et al. (2019), this reduction in variance is of more than academic significance. Large sums of money might (appropriately) turn on it.

We thus take a second approach to measuring the relative performance of specifications. In securities litigation milestones such as class certification or the motion to dismiss or summary judgment stages, courts often require plaintiffs to show that excess returns are statistically significantly different from 0 at levels such as 5% or 10%. We use our simulation evidence to evaluate the performance of various specifications in this task. Let δ be the event-date effect. Modifying (1) to account for this effect yields

[6]We are also working on applying this insight to the many-firm studies commonly used in academic research.

[7]Some machine learning algorithms are more complex to understand and explain. In the interests of keeping our discussion approachable, we exclude consideration of neural nets, which are among the more sophisticated machine learning methods.

[8]As we discuss below, that is not true for specifications that seek to minimize the sum of absolute deviations subject to penalizing constraints. This isn't all that surprising given that we use the MSE metric to evaluate performance, because an objective function targeting least absolute deviations is obviously different from one that targets the sum of squared residuals.

$$r_{it} = g^j(X_{it}) + \delta D_{it} + \zeta_{it}^j, \tag{2}$$

where D_{it} is an indicator variable that equals 1 on an event date and 0 otherwise. (Notice that (1) and (2) are equivalent for non-event dates.)

We consider both the case in which there truly is no event effect, so that $\delta = 0$, and that in which firm value fell on the event date for reasons unrelated to X_{it}, so that $\delta < 0$. Of interest is the result of testing the null hypothesis $H_0 : \delta = 0$ when this null is true (allowing us to evaluate actual test size) and when it is false (allowing us to evaluate actual test power).

Following much practice, we first use the standard approach based on normal critical values.[9] Gelbach et al. (2013) point out that this approach is invalid if excess returns are not truly normally distributed. They show that there are nontrivial consequences of using the standard approach with real-world excess returns, whose non-normality is widely known. Accordingly, we also use the sample quantile (SQ) test proposed by Gelbach et al. (2013). This test works under normality but also is robust to non-normality. Our event-date test results provide several interesting findings.

First, across all specifications, the standard approach under-rejects a true null hypothesis, whereas the SQ test performs almost perfectly across almost all specifications. This is in line with what Gelbach et al. (2013) found using only the simple market model specification, so perhaps it is not surprising. But we use a more recent period than that paper, and we also use a restricted set of firms, and we consider a much more varied set of specifications (including the Fama-French/Carhart factors). Accordingly, our present finding provides additional evidence in favor of the relative superiority of the SQ test over the standard t-test approach.

[9] Analysts often use Student's t critical values instead, because the test statistic has a Student's t distribution under the normality assumption. Because we have a large number of degrees of freedom, the difference between the critical values is negligible for practical purposes.

Second, we find that the specifications that had best performance in the MSE metric also have noticeably better performance in the testing metric (for sufficiently large $|\delta|$, all tests have such high power that the difference is unimportant). Third, this power difference is less than the improvement brought by using the SQ test rather than the standard t-test approach, which can be substantial for intermediate values of δ.

In sum, we find that ML specifications can improve on standard ones. We also find that when testing for statistically significant effects is an analyst's objective, it matters how one tests. Using a method that is robust to non-normality, namely the SQ test rather than the standard approach, improves the performance of ML specifications considerably.

2. Prior Literature

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 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.[10] 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

[10] Kothari and Warner (2007) report that over 500 papers containing event studies were published betwen 1974 and 2000 in just the top five finance journals.

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. They showed that stock returns departed from normality, but still they found OLS-based methods to be largely robust to parametric concerns in applications of interest.

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 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 in important situations, the Type I error rate will be larger than the nominal level that holds when the assumption of normality is correct. This is particularly true for stocks with high kurtosis (Hein and Westfall, 2004), which is not surprising given the departure from normality entailed by this distributional feature. 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 another modification, which they termed the sample quantile (SQ) test. To perform a lower-tailed version of this test with classical significance level α, one ranks the estimated excess returns from the market model regression and determines

whether the event-date excess return is more negative than the α-quantile of the empirical distribution of estimated excess returns from the pre-event window.[11] Using a dataset containing the returns for all securities in the Center for Research in Security Performance's (CRSP) database from 2000 to 2007, Gelbach et al. (2013) 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 account for the true variance of market model excess returns. 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 then be applied.[12] 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

[11] 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.

[12] We do not implement the GARCH approach in this paper, but presumably one could do so.

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).[13] 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 significant determination is in the method used to characterize 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 apply 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

[13]For an interesting example of a case in which the speed of capitalization was at issue, see *In re Apollo, Inc., Securities Litigation.* ADD MORE.

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).

Recall from (1) that we write expected returns for specification k as $r_{it} = g^k(X_{it}) + \zeta_{it}^k$, where r_{it} is the measure of the daily stock return, g^k is some function that captures the details of specification k, and ζ_{it}^k is the excess return under that specification on date t for firm i. If we assume that the excess return has mean 0, then the expected return is $E[r_{it}] = g^k(X_{it})$. As an example, the simple market model has an expectation that is an affine function of the daily market return, so

$$g^{MM}(\cdot) = \alpha^{MM} + M_t\beta^{MM}.$$

The associated excess return is

$$\zeta_{it}^{MM} = r_{it} - \alpha^{MM} + M_t\beta^{MM}. \tag{3}$$

When we view an event study as a prediction problem, our goal is to isolate the portion of the return r_{it} that cannot be explained with available variables X_{it}. That is of course true of the conventional least squares approaches as well, but ML differs by allowing more scope for data-driven selection of which variables are ultimately included, and how. One way to understand ML methods is that they use more flexible g functions for the expected return. Another is to think of them as starting with familiar expected return functions and then using certain nonlinear alterations to the objective function. Regardless of which view one takes, the end result is an estimator that differs from conventional least squares estimators in its use of available data.

Below we consider a total of 32 specifications, each simulated 10,000 times using 250 estimation-set observations on a firm's daily returns and one out-of-sample "event date" return. The firms and event dates are randomly selected, so that event date actual and

excess returns are not systematically related to anything about our specification choices. The exact ways we pick the 250 dates vary a bit with specification, as described in the discussion below.

Of the 32 specifications, 16 include the Fama-French and Carhart factors. The other 16, including the simple market model specification described just above, do not.

Write $it(b)$ to indicate the firm-date used as the event-date for simulation replicate $b \in \{1, 2, \ldots, 10000\}$, so that the estimated event-date excess return for specification k is $\widehat{\zeta}^k_{it(b)}$. At present we take four approaches to comparing performance across models–two involving variance of the estimated event-date excess return, and two involving the performance of significance tests based on event-date excess return.[14]We discuss details of these approaches below, just before we report the corresponding empirical results.

3.1. Specifications Used

As noted, we consider 32 specifications. There are 16 distinct approaches to estimation, and for each of these we consider two specifications: one that includes the four Fama-French and Carhart (FFC) variables, and one that does not. Here we describe the 16 distinct specifications.

For reference, Table 3.1 provides a list of detail on the specifications we discuss below; the table includes columns with the specification acronym and number, as well as information about the included explanatory variables, the objective function used, and the algorithm on which it is based.

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.

[14]We may add another approach in future versions, which would assess how the different specifications perform at damages estimation in those situations in which the excess return would be found statistically significant in securities litigation. This would allow us to address points made about bias in damages estimation by Brav and Heaton (2015) and Dove et al. (2019)

Table 1: List of Specification Acronyms, Numbers, and Descriptions

Acronym	Number	Explanatory Variables	Q Minimization Target	Algorithm
MM	1	BM*	Squared-error	Standard^{+} OLS
MMPI	2	BM*, EWPI*	Squared-error	Standard^{+} OLS
Med	3	BM*, EWPI*	Squared-error	Standard^{+} quantile regression
GR	4	BM*, EWPI*	Absolute deviation	Standard^{+} quantile regression
Trimean	5	BM*, EWPI*	Absolute deviation	Standard^{+} quantile regression
ENR	6	BM*, EWPI*; Model-generated†	Squared-error	Elastic net regularization
ENR-U	7	BM*, UPI*; Model-generated‡	Squared-error	Elastic net regularization
ENR-FMI	8	BM^{*,FMI}, EWPI*; Model-generated^{FMI}	Squared-error	Elastic net regularization
ENR-LEW	9	BM*, LEWPI*; Model-generated$^{\diamond}$	Squared-error	Elastic net regularization
ENR-Med	10	BM*, EWPI*; Model-generated†	Absolute deviation	Elastic net regularization
ENR-U-Med	11	BM*, UPI*; Model-generated‡	Absolute deviation	Elastic net regularization
ENR-FMI-U-Med	12	BM^{*,FMI}, UPI*; Model-generated^{FMI}	Absolute deviation	Elastic net regularization
LLRF-2F	13	Model-generated†	Squared-error	Random forest
LLRF-U	14	Model-generated‡	Squared-error	Random forest
ENR-TSCV-2F	15	BM^{*,FMI}, EWPI*; Model-generated†	Squared-error	Elastic net regularization
ENR-TSCV-U	16	BM^{*,FMI}, UPI*; Model-generated‡	Squared-error	Elastic net regularization

* The following abbreviations are used for variables (a/k/a factors):

- "BM" refers to the broad market index.
- "EWPI" refers to the equally weighted peer index.
- "UPI" refers to the unrestricted peer firms included, based on elastic net regularization estimates.
- "LEWPI" refers to the lasso selection-based equally weighted peer index.

$^{+}$ "Standard" means estimated as usual, with no regularization.

† The BM and EWPI variables are available for selection via the estimation algorithm.

‡ All peer firms are available for selection via the estimation algorithm.

$^{\diamond}$ All peer firms are available for selection via first-step lasso and then used to create an equally-weighted index that is used in the final estimation step of the algorithm.

^{FMI} The daily return series for the broad market index is forced to be included in the final estimation step, so that if regularization otherwise would eliminate this variable, it is prevented from doing so.

Here we use the return on the S&P 500 Index as a proxy for aggregate movement in the stock market. The specification for the 250-day estimation window is:

$$r_{it} = \alpha^{MM} + \beta^{MM} mkt_{it} + \zeta_{it}^{MM}, \tag{4}$$

with $E[\zeta_{it}^{MM}] = 0$ given the presence of the constant and the fact that the parameters will be estimated using OLS.[15] For date 251, the estimated (equivalently, predicted) 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 (4).

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 daily return index from firms in the same SIC industry as firm i, which we call $peer_{it}$. This kind of peer index is commonly used in litigation. We construct our version using all firms in the same 4-digit SIC industry as firm i, unless there are fewer than eight such firms, in which case we use all firms in the same 3-digit SIC industry as i.[16] The return specification is:

$$r_{it} = \alpha^{MMPI} + \beta_1^{MMPI} mkt_{it} + \beta_2^{MMPI} peer_{it} + \zeta_{it}^{MMPI}, \tag{5}$$

and the excess returns are estimated as

$$\widehat{\zeta}_{it}^{MMPI} \equiv r_{i,251} - [\hat{\alpha}^{MMPI} + \hat{\beta}_1^{MMPI} \times mkt_{i,251} + \hat{\beta}_2^{MMPI} \times peer_{i,251}],$$

where $(\hat{\alpha}^{MMPI}, \hat{\beta}_1^{MMPI}, \hat{\beta}_2^{MMPI})$ is the vector of OLS estimates of the coefficients in equation (5).

[15] We do not impose or assume the stronger assumption, $E[\zeta_{it}^{MM} | mkt_{it}] = 0$. If this assumption did hold, then the parameter β^{MM} could be understood as a causal effect; without the assumption, β^{MM} is merely a linear projection parameter. See Chapter 2 of Wooldridge (2002) on these matters.

[16] If there are fewer than five such firms we drop them from consideration.

3.1.3. Specification 3 - Median Regression (Med)

Koenker and Bassett (1978) show that estimators corresponding to the minimization of mean-squared error can perform poorly when the residual distribution is sufficiently long-tailed. It is now well established that stock returns are non-normal distribution and have excess mass in the tails of the distribution (see, e.g. Gelbach et al. (2013) and references therein). Accordingly, we consider various quantile regression-based estimators, because Koenker and Bassett (1978) show that these can be robust to non-normality.

Specification 3 uses median regression to estimate the coefficients in specification 2. Median regression can be understood as choosing b to minimize $\sum_{t=1}^{250} |\zeta_{it}^{Med}(b)|$.[17] The estimated excess return for specification 3 is $\widehat{\zeta}_{it}^{Med}(b) \equiv r_{i,251} - [\hat{\alpha}^{Med} + \hat{\beta}_1^{Med} \times mkt_{i,251} + \hat{\beta}_2^{Med} \times peer_{i,251}]$, 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 a proposal, analyzed in the univariate context, by Gastwirth (1966)[18] 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 τ-quantile regression coefficients for each choice of $\tau \in \{1/3, 1/2, 2/3\}$. Following Gastwirth, we then calculate the weighted average

$$\hat{\beta}^{GR} \equiv 0.3\hat{\beta}\left(\frac{1}{3}\right) + 0.4\hat{\beta}\left(\frac{1}{2}\right) + 0.3\hat{\beta}\left(\frac{2}{3}\right)$$

. The estimated excess returns in the Gastwirth specification is $\widehat{\zeta}_{it}^{GR}(b) \equiv r_{it} - X_{it}'\hat{\beta}^{GR}$.

[17] Equivalently, it chooses b to minimize $\sum_{t=1}^{250} \rho_{0.5}(\zeta_{it}^{Med}(b))$, where ρ_τ is the check function defined so that $\rho_\tau(\zeta) = \zeta[\tau - 1(\zeta < 0)]$.

[18] See Joseph L. Gastwirth (1966), "On Robust Procedures", *Journal of the American Statistical Association*, 61:316, 929-948, DOI: 10.1080/01621459.1966.10482185.

3.1.5. Specification 5 - Trimean Regression (Trimean)

This is the final quantile regression specification based on Koenker and Bassett (1978). It is similar to the Gastwirth specification, Specification 4, but it uses different weights and quantile regression estimators from different points in the $[0, 1]$ interval. Instead of using the 1/3-, median-, and 2/3-quantile regression estimates, the Trimean specification uses a weighted average of the 1/4-, median-, and 3/4-quantiles. The estimated coefficient vector is

$$\hat{\beta}^{Trimean} \equiv .25[\hat{\beta}\left(\frac{1}{4}\right) + 2\hat{\beta}\left(\frac{1}{2}\right) + \hat{\beta}\left(\frac{3}{4}\right)].$$

The estimated excess return for the Trimean specification is $\widehat{\zeta}_{it}^{Trimean}(b) \equiv r_{it} - X_{it}'\hat{\beta}^{Trimean}$.

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.

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

$$Q(c; a, \lambda) \equiv \zeta_i(c)'\zeta_i(c) + \lambda\left(\frac{1-a}{2}c'c + a\sum_{j=1}^{p}|c_j|\right), \tag{6}$$

where a and λ 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 specification mean squared error, we do the following for each set of 251 observations:

- Randomly group the data into ten random groupings, known in the ML literature as folds.
- For each of the nine values of $a \in \{0.1, 0.2, \ldots, 1\}$ use the ten randomly created folds with a procedure known as cross-validation, to find the minimizing value of (c', λ); call the resulting estimates, $c^*(a)$ and $\lambda^*(a)$.
- Denote as a^* the value of a that yields the lowest MSE among the nine MSE-minimizing values.
- Set $\hat{\beta}^{ENR}$ equal to the estimates of the coefficients c with $a = a^*$ and $\lambda = \lambda^*(a^*)$, i.e., $\hat{\beta}^{ENR} = c^*(a^*)$.
- The estimated excess returns are calculated as usual, i.e., as $\widehat{\zeta}_{it}^{ENR}(b) \equiv r_{it} - \hat{\beta}^{ENR} X_{it}$.

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 specification 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_{peer} peer firms' returns equal to N_{peer}^{-1}. Thus specification 7 is more flexible specification in terms of index creation than all the any

specification whose regressor set is the market return and equally weighted peer index.[19]

For specification 7, we implement the unconstrained peer firm returns specification 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 $2 + N_{peer}$ dimensions—one for a, one for c_1, and one for each of the N_{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 daily return of the market index variable. Depending on the data, specification 7's algorithm might drop the market index regressor before calculating coefficient estimates. Specification 8 differs from specification 7 only in that this is not allowed: the final estimates are based on an estimation step that includes the daily return of the market index in the regressor set. We investigate this specification out of a belief that some experts and courts might insist that the market index be part of the specification 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 (ENR-LEW)

The first step of specification 9 can be thought of as a version of specification 6's elastic net regularization, with a set to 1. Once we have a set of selected peer firms we use them to calculate an equally weighted peer index. This differs from specification 6 because in that specification we use the estimated coefficients to weight peer firms' daily returns, whereas in specification 9 we use OLS to estimate coefficients on the market return and an equally-

[19]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, because it allows for more covariates than observations; it does so by dropping weakly correlated controls from the estimation equation. This is one reason lasso is frequently used for model selection problems.

weighted peer-firm index whose member firms are constructed based on first-step estimated lasso coefficients.[20]

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 equally-weighted 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 first part of equation (6)'s objective function Q 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. Thus the objective function is

$$Q^{2FM-Med}(c;a,\lambda) \equiv \sum_{j=1}^{p}|c_j| \qquad \lambda\left(\frac{1-a}{2}c'c + a\sum_{j=1}^{p}|c_j|\right). \tag{7}$$

All else proceeds as in the discussion in specification 6.

3.1.11. Specification 11 - Median Regularization with Market Index and All Peer Firms (ENR-U-Med)

This specification changes specification 10 in the same way that specification 7 changed specification 6. Like specification 10, it uses an objective function that minimizes the sum of absolute deviations and has regularization built in. Like specification 7, it allows peer firms' coefficients to be determined by the data rather than using an equally-weighted average.

[20]First we use lasso to estimate coefficients for peer firms. We then construct the equally weighted peer index based on those peer firms that have nonzero lasso coefficients. Then we run OLS, whose estimated coefficients are the final ones we use to estimate excess returns. In any simulation replication on which there are no peer firms with nonzero lasso coefficients, the last step uses OLS estimation of the one-factor market model.

3.1.12. Specification 12 - Median Regularization with Forced Market Index Inclusion and All Peer Firms (ENR-FMI-U-Med)

This specification changes specification 11 in the same way that specification 8 changed specification 7: it forces the inclusion of the daily return of the broad market index in the final model. Otherwise this specification is computed like the one in specification 11.

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. This is in part because they require little model tuning, so that their out-of-the-box performance is superior compared to more complex machine learning methods such as neural nets. Random forests are forms of regression trees that are notably effective when used with large numbers of "features"–explanatory variables, in econometric parlance–that are not truly related to the outcome variable, as is likely the case in event studies.[21] Here we use the local linear random forest method from Friedberg, Tibshirani, Athey, and Wager (2018), who point out that we can view random forests as an adaptive kernel method. Pairing the random forest-generated kernel with local linear regression adjustment is desirable when the relationship of interest is smooth, as in a regression model of the form typically estimated with event studies. We refer interested readers to Friedberg et al. (2018) for details.

Specification 13 uses local linear random forests with two factors: a market and peer index.

3.1.14. Specification 14 - Local Linear Random Forest Regressions With Market and Peer Returns (LLRF-U)

This specification 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.

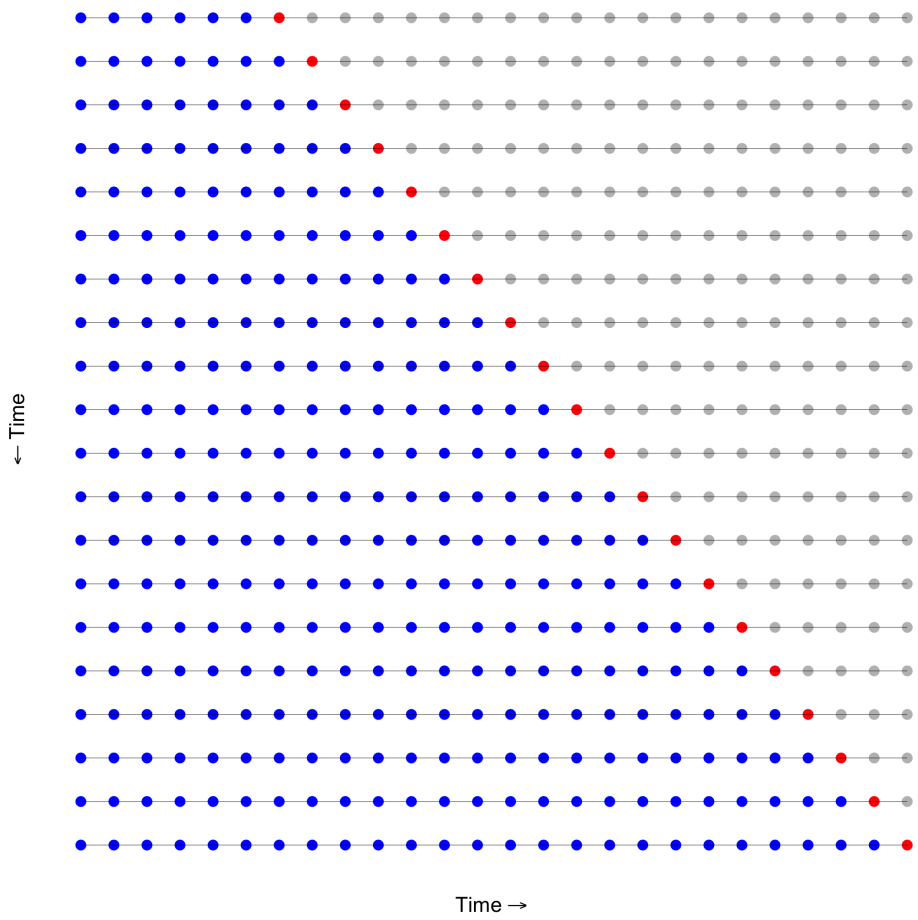
[21] Of course if we knew which variables were unrelated we would simply exclude them from estimation; the challenge to which ML methods are meant to rise is the need to use the data itself to determine which variables are unrelated to the outcome of interest.

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

The penalized regression approaches described above estimate the penalization parameters α and λ 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). Prediction 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 specification 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.

4. Simulation 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.

4.1. Comparison Approach 1: Excess return variance normalized against within-date in-sample variance of the simple market model

The squared value of the excess return for specification k and firm-date $it(b)$ is $\widehat{w}^k_{it(b)} \equiv (\widehat{\zeta}^k_{it(b)})^2$. Using the convention that the event date is labeled $t = 251$, the in-sample prediction of interest for specification k on simulation replicate b is $\widehat{w}^k_{i251(b)}$. The estimated in-sample mean squared error (MSE) for specification k is

$$\widehat{MSE}^k_{oos} \equiv \frac{1}{250} \sum_{i=1}^{250} \widehat{w}^k_{i251(b)}. \tag{8}$$

Our first comparison approach is meant to deal with heteroskedasticity in excess returns. There are some dates on which important events really did occur, and for unmodeled reasons. Excess returns will be especially large on such days.[22] One might worry that this phenomenon will cause $\widehat{MSE}^k_{oos}$ to be unduly sensitive to a relatively small number of especially high-variance dates. We address this concern by using a metric that normalizes within-date according to that date's in-sample MSE for the simple market model. One approach to such a normalization would be to compute the average, across the 10000 simulation replicates, of the ratio of $\widehat{w}^k_{i251(b)}$ to $\widehat{w}^{MM}_{i251(b)}$. That should account for both in-sample (the 250 non-event

[22]Similarly, some estimation windows will encompass real events, which will tend to cause the specifications we estimate to have especially poor out-of-sample fit. That will exhibit in the form of apparently large squared residuals on our $t = 251$ days.

dates) and out-of-sample (the randomly selected event date) heteroskedasticity. But it also runs a risk. On an event date when the (normalizing) market model happens to predict almost perfectly due only to estimation error, rather than model signal, the denominator of $\widehat{w}^k_{i251(b)}/\widehat{w}^{MM}_{i251(b)}$ will be close to 0, causing the overall ratio to be extremely large. This effect could dominate our normalized metric of performance, obscuring its signal with estimation-induced noise. To avoid this kind of effect, we instead use the following metric for each model k:[23]

$$\widehat{R}^k_{het} \equiv \frac{1}{10000} \sum_{b=1}^{10000} \left(\frac{\widehat{w}^k_{i251(b)}}{\widehat{MSE}^{MM}_{est(b)}} \right), \tag{9}$$

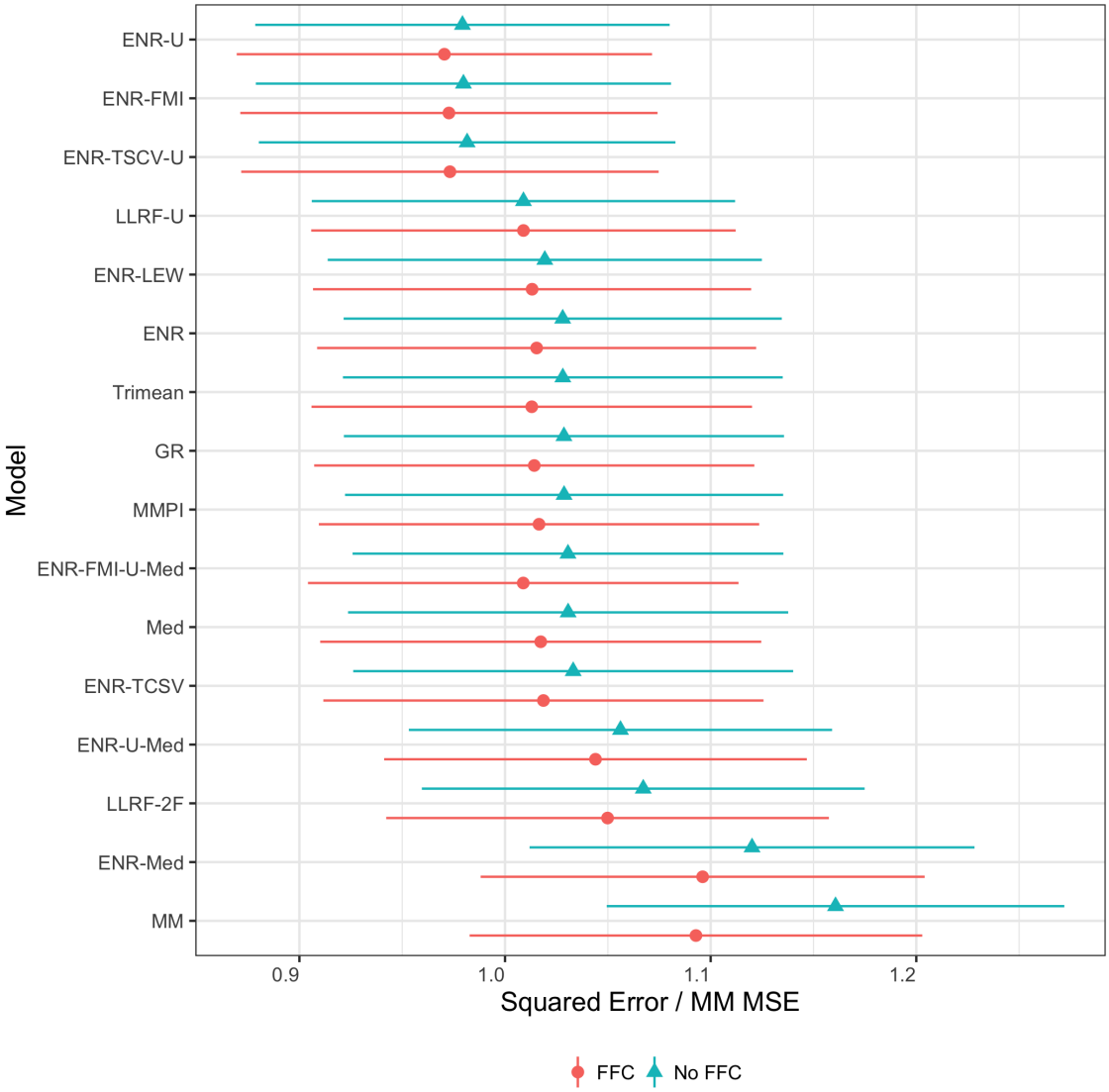
where $\widehat{MSE}^{MM}_{est(b)} \equiv \frac{1}{250} \sum_{i=1}^{250} \widehat{w}_{it(b)}$ is the average squared estimated excess returns over the 250-day estimation window used in simulation replicate b for the Market Model specification.

Thus, the $\widehat{R}^k_{het}$ metric normalizes the squared event date excess return by the *in*-sample estimate of the MSE for the simple market model (i.e., the specification that includes only a constant and the daily market return). Because this denominator is constructed by averaging over a large sample of dates, it does not engender the signal:noise problem described just above.

Figure 1 plots the mean of the normalized prediction errors of the 16 models with (circles) and without (triangles) the Fama-French Carhart factors, together with 95% confidence intervals. The order in which the specifications are listed on the vertical axis is determined by performance in the specifications without these factors, so that the specification reported in the top row is the one with lowest value of $\widehat{R}^k_{het}$, and the one that appears in the bottom row is the one with the greatest value.

[23]Note that the subscript "het" on $\widehat{R}^k_{het}$ refers to the concern about heteroskedasticity whose possibility motivates this metric.

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, i.e., $\widehat{R}^k_{het}$, according to which each event-date's squared prediction error is normalized by the mean squared error of the estimation-period residuals from the simple market model. We plot the estimates both with (FFC) and without (No FFC) Fama/French/Carhart Factors.

Figure 1 shows that the worst-performing specification without the FFC factors is the simple market model. Notice that its value of $\widehat{R}^k_{het}$ exceeds 1: it is roughly 1.16, which means the variance of estimated event-date excess returns for the simple market model is about

16% greater than the in-sample variance for that same model. This difference between the in- and out-of-sample variance illustrates the empirical importance of overfitting.

The best-performing specification, both with and without the FFC factors, is ENR-U (specification 7). Recall that this specification uses penalization, targets the squared value of the residual (i.e., variance), and allows the estimation algorithm to select the coefficients on the broad market index as well as on each peer firm. This is the most flexible of the variance-targeting penalized regression specifications we considered, so it is perhaps not surprising that it performs best. For ENR-U, the out-of-sample event-date excess return variance is roughly 97-98% of the in-sample variance of the simple market model.

Two other specifications are so close to ENR-U as to be essentially indistinguishable. The ENR-FMI specification (specification 8) differs from ENR-U only in that it forces the inclusion of the broad market index in the final estimation. The TSCV-U specification differs from ENR-U only in that it uses a more dynamically robust method of cross-validation to select variables and estimate coefficients.

The next 9 specifications ranked below the top triumvirate perform in relatively indistinguishably ways. All have out-of-sample average normalized event-date excess return variance between 100% and 104% of the in-sample variance of the simple market model, with or without the FFC factors included. The next 2 specifications, ENR-U-Med and LLRF-2F, have average normalized event-date excess return variance between 104% and 107% of the in-sample variance for the simple market model without the FFC factors. The two worst specifications are the simple market model and the same model estimated via median regression–specification "Med," which is specification 3. These specifications have average normalized variance of about 109% with the FFC factors, and 112% and 116% without them.

We can draw a number of lessons from Figure 1. One is that regularization–which is to say, using ML algorithms–does appear to reduce event-date excess return variance. This conclusion should be tempered in two ways. First is that the 95% confidence intervals in Figure 1 are quite wide, suggesting that some of the reduction might be due to sampling and

simulation noise. That said, the different $\widehat{R}^k_{het}$ numbers plotted in the figure are obviously highly positively correlated across specifications, so it is difficult to draw meaningful inferences based on model-specific confidence intervals. A future task for us is to appropriately test whether the different values plotted in the figure are statistically significantly different from each other. This task is complicated not only by their positive correlation, but also by the obvious multiple-inferences issue that arises in such a task.

A second lesson from Figure 1 is that although including the FFC factors makes a comparatively large difference for the simple market model (MM), doing so generally seems to be less important than including some sort of peer-firm adjustment. To see this, consider the MMPI specification, which uses standard OLS estimation in a specification that includes only the broad market index and the equally-weighted peer index. This specification does noticeably better *without* the FFC factors than the simple market model does *with* them. The same is true for all the specifications that include some sort of peer-firm adjustment, with the exception of the ENR-Med specification.

Third, among the regularized specifications, the three best performers target squared error. Fourth, these specifications either are forced to include the broad market index or are allowed to estimate the role of peer firms in an unrestricted manner (either of these is sufficient to perform better than the base ENR specification). Fifth, it does not appear that time-series cross-validation is per se important: although the ENR-TSCV-U specification is in the top group, the ENR-TSCV-2F specification is not; it performs about the same as several other estimators that use either standard cross-validation or don't use regularization at all–e.g., the MMPI, Med, GR, and Trimean specifications, all of which use unpenalized estimation algorithms but include both the market model and the equally-weighted peer index.

In sum, the best performance in Figure 1 comes when regularization is paired with some sort of peer-firm adjustment and the FFC factors.

4.2. Comparison Approach 2: Excess return variance normalized against the simple market model's average event-date variance

Our second approach to comparing model performance is to compute the ratio of estimated out-of-sample MSE for model k to the same variable for the simple market model. That is, we compute:

$$\widehat{R}^k_{oos} \equiv \frac{\widehat{MSE}^k_{oos}}{\widehat{MSE}^{MM}_{oos}}. \tag{10}$$

Notice that $\widehat{R}^k_{oos}$ and $\widehat{R}^k_{het}$ differ: $\widehat{R}^k_{het}$ is an average of a ratio, whereas its counterpart $\widehat{R}^k_{oos}$ is a ratio of averages. Whereas $\widehat{R}^k_{het}$ normalizes within dates, $\widehat{R}^k_{oos}$ instead averages across dates and only then normalizes. The $\widehat{R}^k_{oos}$ approach would be problematic if there is so much heteroskedasticity in event-date excess returns that the variance from a small share of firm-date pairs dominates the average of overall variances. As long as that is not the case, $\widehat{R}^k_{oos}$ will be a meaningful measure of performance.[24] Finally, we note that by construction $\widehat{R}^{MM}_{oos} = 1$, so that other specifications' values of $\widehat{R}^k_{oos}$ may be regarded in terms of the percentage reduction in out-of-sample variance they achieve by comparison to the market model.[25]

Figure 2 plots the various specifications' values of $\widehat{R}^k_{oos}$ in the same order as the specifications' values of $\widehat{R}^k_{het}$ were plotted in 1. Although we believe no constraint forces the specifications to perform similarly with $\widehat{R}^k_{oos}$ as with $\widehat{R}^k_{het}$, Figure 2 indicates that, broadly considered, they do.

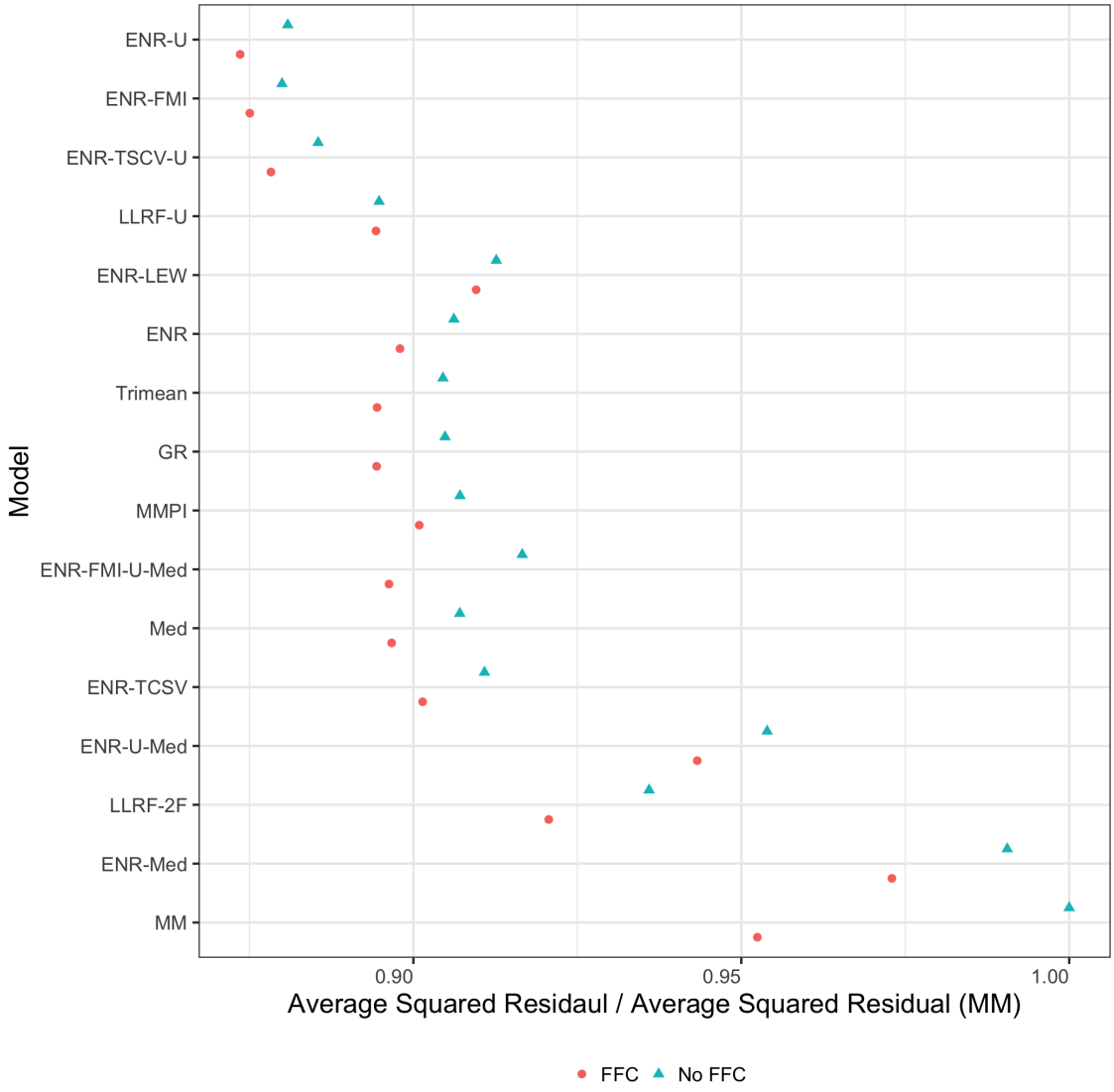
Most notably, the three best-performing specifications again are ENR-U, ENR-FMI, and ENR-TSCV-U. When the FFC factors are included, all three have average event-date excess

[24] Notice that $\widehat{R}^k_{oos}$ is not simply $\widehat{R}^k_{het}/\widehat{R}^{MM}_{het}$, because each $\widehat{R}^k_{het}$ is an average of ratios, whose denominator, $\widehat{MSE}^{MM}_{est(b)}$, changes across simulation replicates indexed by b.

[25] As we will see momentarily, none has greater event-date excess return variance than the simple market model without the FFC factors.

return variance equal to roughly 85% of the variance for the simple market model without the FFC factors. Without the FFC factors, each of the three specifications performs worse by 1-2 percentage points.

Figure 2. Ratios of Average Squared Residual to MM



Note: Figure 5 plots the average value of $\widehat{R}^k_{oos}$ across specifications; this is the average squared residual for each model divided by the average squared residual for the simple market model (MM). The models are reported in order of their predictive power in the no-FFC models, i.e., in the same order as in Figure 1.

The other specifications perform in qualitatively similar ways as when we used the first comparison approach. The most notable difference is that the ENR-LEW specification does a good bit worse than the other eight specifications in the second-best performance group, whereas in Figure 1 it was (marginally) the best of these 9 specifications.

All in all, though, the results in Figure 2 leave us with the same basic conclusions we had after viewing Figure 1.

4.3. Comparison Approach 3: Significance test performance using the standard parametric testing approach

We have seen that regularization can enhance the precision of event study excess return estimates. We now assess how important precision improvements are for significance tests of whether excess returns are significantly different from 0. These tests are important in securities litigation, because class certification and resoultion of motions to dismiss or for summary judgment may turn on their results.

Comparison approach 3 considers both the Type I error rate, also known as size, and the power (one minus the Type II error rate) of the standard approach of comparing t-statistics to critical values based on the standard normal distribution.[26] To assess actual size with a nominal size-α test, on each simulation replicate b we "reject" the null hypothesis of no event effect whenever the ratio of the estimated event-date excess return to its estimated standard deviation is less than the α-quantile of the standard normal distribution:

$$\widehat{T}_b^k \equiv \frac{\widehat{\zeta}_{i251(b)}^k}{RMSE_b^k} < z_\alpha, \tag{11}$$

where $RMSE_b^k \equiv \sqrt{\frac{\sum_{t=1}^{250}(\zeta_{bt}^k)^2}{250}}$ is the in-sample estimate of the standard deviation for the

[26] Adjustments for degrees of freedom are functionally irrelevant given the numbers of degrees of freedom we have with 250 estimation dates.

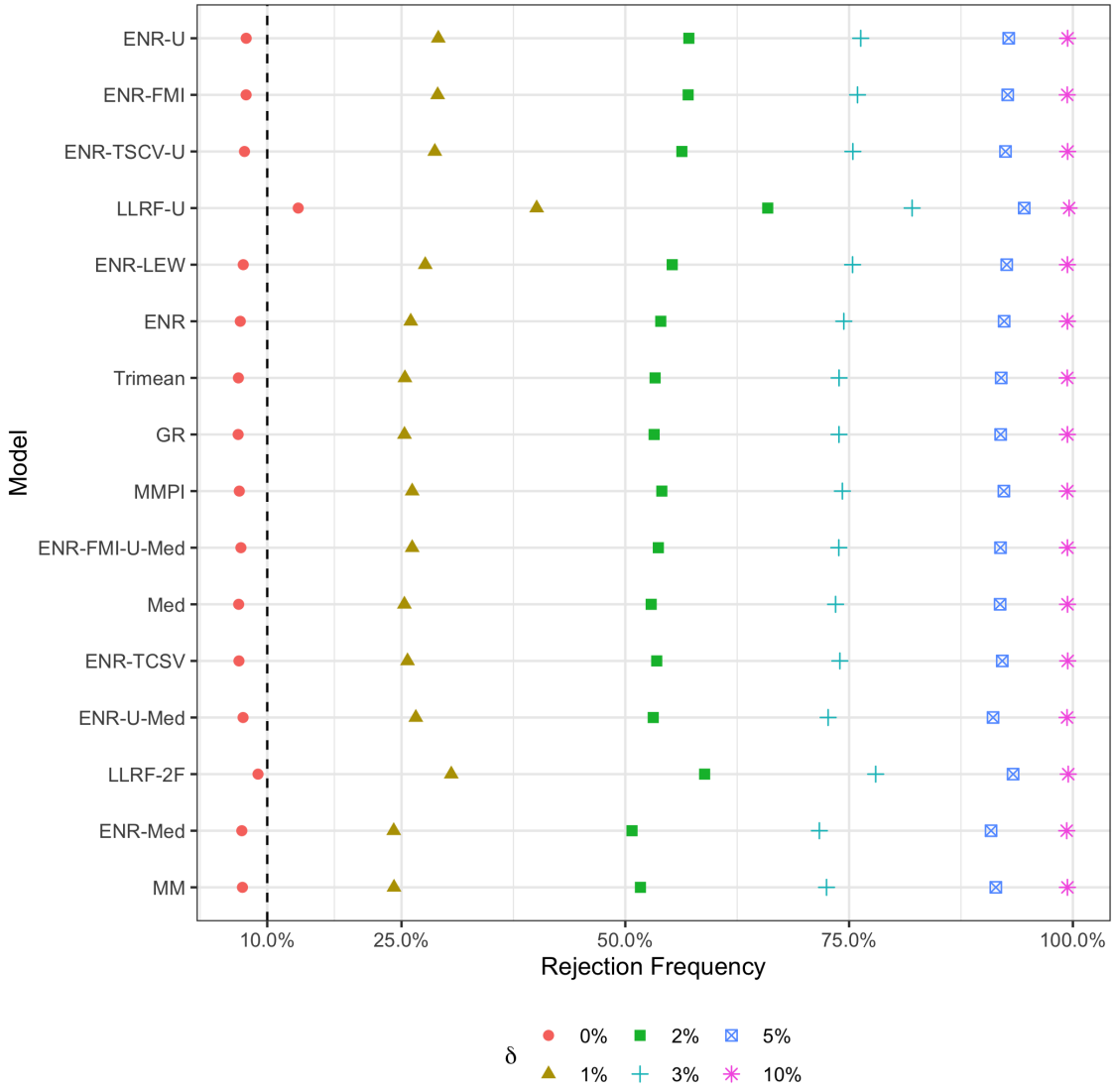
event-date excess return on simulation replicate b. We then compute the share of our 10,000 simulation replicates on which this test rejects. That share is the estimated size (equivalently, Type I error rate) of specification k for a nominal size-α test.

To assess power, we must adjust $\widehat{T}_b^k$ to account for the "true" magnitude of the event effect that is of interest. We assume that events of interest cause firm value to fall by the amount δ, so that adjusted event-date returns are $r_{i251}^{\delta} = r_{it} - \delta$. We consider drops of magnitude $\delta \in \{.01, .02, .03, .05, .10\}$; given our use of logged returns, this means we investigate power against events that cause firm value to fall by approximately 1%, 2%, 3%, 5%, and 10%. The adjusted event-date excess return is thus $\widehat{\zeta}_{i251(b)}^{k,\delta} = \widehat{\zeta}_{i251(b)}^{k} - \delta$, and for our power analyses we replace $\widehat{\zeta}_{i251(b)}^{k}$ with $\widehat{\zeta}_{i251(b)}^{k,\delta}$ in the test condition in 11. Because the critical value on the right hand side of that condition is fixed, the estimated rejection rate will increase with the assumed magnitude of the event effect. Finally, we note that size may be thought of the rejection rate when $\delta = 0$.

Figure 3 reports simulation results for the percentage of simulation replicates on which these tests reject, using a significance level of $\alpha = 0.10$ (so that $z_\alpha = -1.28$), and considering only the 16 specifications with the FFC factors included. The dashed vertical line at 10% is the nominal size of the test: when $\delta = 0$, the null hypothesis of zero event effect is correct, and a test with correct size would reject exactly 10% of the time. Instead, almost all of the specifications reject considerably less often than that–only about 7-8% of the time.[27] These findings as to substantial size distortions echo those in Gelbach et al. (2013).

[27]The exceptions are the two specifications that use random forest-based local linear regression. The version that allows an unrestricted peer index actually over-rejects almost as often as most of the specifications under-reject; the version that uses the equally-weighted peer index rejects just below the nominal rate of 10%.

Figure 3. Power Analysis - Standard Approach Tests of Significance



Note: Figure 3 plots the rejection frequencies for our 16 models, estimated with the inclusion of Fama-French/Carhart factors. The parameter δ is the level of the event effect. Rejection is based on the standard approach–comparing t-statistics to a standard normal critical value.

For values of δ above zero–i.e., when there really was an event effect–there is some variation in performance across specifications. The two best-performing specifications are those that use random forest-based local linear regression, which is not surprising given their greater Type I error rates. More notable, perhaps, is the rightward drift of the other

models' rejection rates when the true event effect is a drop in firm value of 1% (triangles), 2% (squares), or 3% (plus-signs). The specifications' rejection rates are reported in the same order they were in Figure 1, so this rightward lean indicates that specifications with lower excess return variance according to the $\widehat{R}^k_{het}$ metric tend to have higher power for small to moderate event-effect sizes. These differences might be practically significant in real-world litigation, although investigating that question directly is beyond the scope of the present paper.

4.4. Comparison Approach 4: Significance test performance using the sample quantile test

The poor size of the standard approach tests exhibited in Figure 3 is unsurprising given (i) the well known non-normality of excess returns and (ii) the arguments made and evidence provided in Gelbach et al. (2013). That paper shows that when excess returns are non-normal, the standard approach – t-tests using critical values based on the standard normal distribution – may lead to serious size distortions like those we see in Figure 3.

Gelbach et al. (2013) propose an alternative, based on the sample quantiles of the empirical distribution function (EDF) of estimated excess returns from the estimation window. They term their test the SQ test, and they show that as the number of dates in the estimation window grows, the SQ test's size converges to the nominal level. Thus the SQ test has asymptotically correct size, where the asymptotics in question have to do with the estimation window length.

Although Gelbach et al. assumed that the return specification they used was correct, introspection shows that that assumption is unnecessary for the SQ test to have correct size. A simple informal argument will suffice for present purposes. As long as the event date is just like estimation-window dates but for the presence of an additive event effect, event-date excess returns based on a fixed specification k will come from the same data generating process as estimation-window returns, up to a location difference due to the event effect–

which is zero under the null hypothesis anyway. The Glivenko-Cantelli theorem then implies that the EDF of estimation-window excess returns is a consistent estimator for the true distribution function of the event-date excess return. Accordingly, the sample quantiles of the estimation period are consistent estimators for the true quantiles of the event-date excess return under the null hypothesis. And that means that the sample α-quantile may be used as a critical value for testing the null hypothesis of zero event effect. Because nothing about this argument requires the specification in question to be correct, the SQ test should have asymptotically correct size for each of the specifications we investigate here.

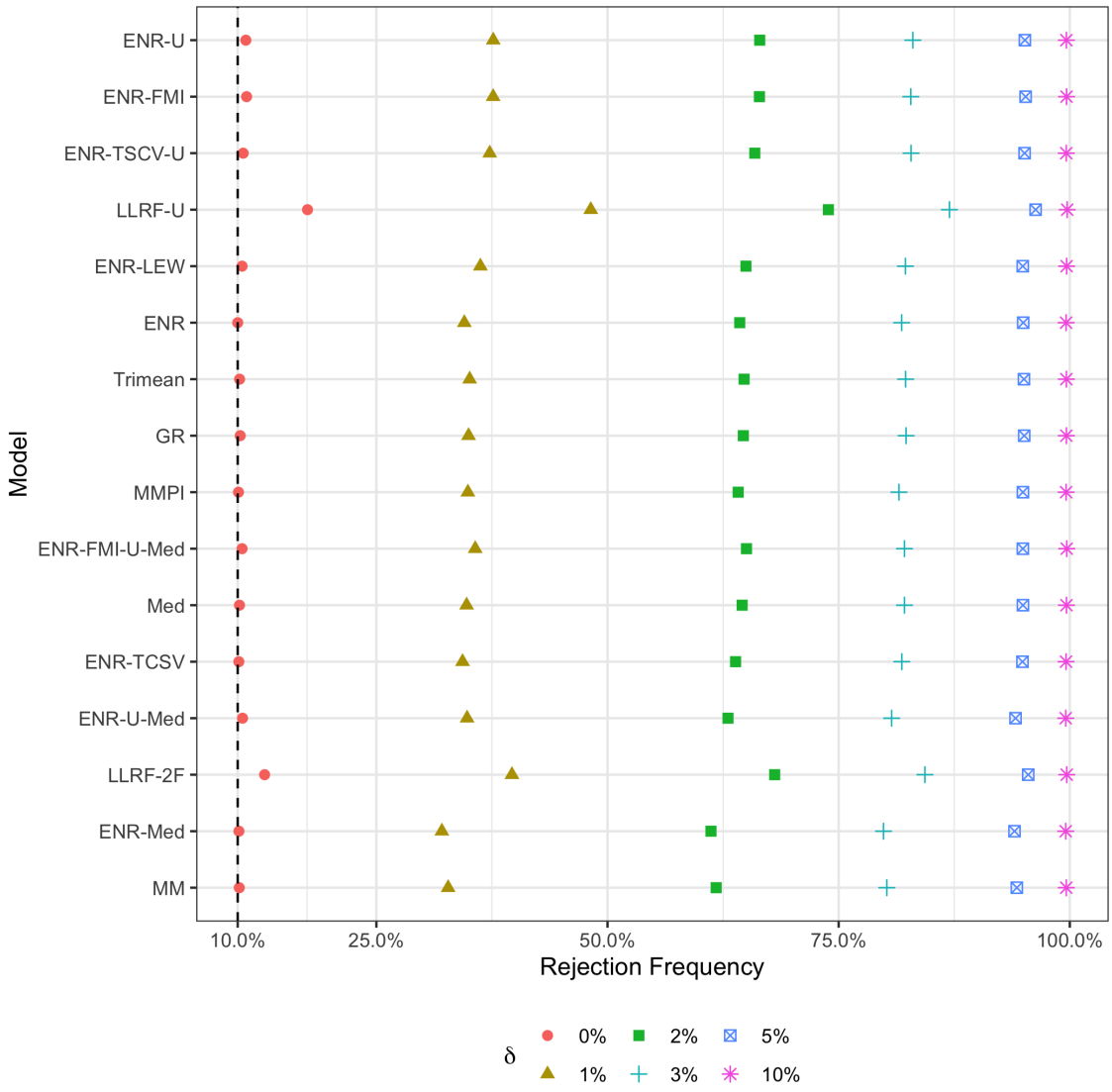
Figure 4 reports SQ test rejection rates for the same values of δ investigated using the standard approach tests reported in Figure 3. As with the standard approach tests, we use a nominal test size of $\alpha = 0.10$. This is implemented in the SQ test by comparing the event-date excess return on each simulation replicate to a critical value that equals the 25^{th} most negative estimated excess return, because that value is the sample 0.10-quantile of excess returns.

There are two specifications whose Type I error rates depart noticeably from 10%. Each involves the random forest-based local linear regression specifications. We believe this may be due to a glitch in the R code implementing the optimization algorithm we used to implement these specifications. We are still investigating that issue. (These two specifications also have substantially greater rejection rates for values of δ above 0, which is to be expected in light of their distorted size; we will not discuss them further in this draft.)

Otherwise, as expected, the figure shows that the Type I error percentages are virtually identical to the nominal level of 10%. Not surprisingly, given the downward size distortions of the standard approach, the power performance of the SQ test is also considerably better than that of the SQ test. For example, whereas the standard approach led to rejection percentages clustered around 25% when the true event effect was a drop in firm value of about 1%, for the SQ test power clusters roughly around 35%. Power is noticeably elevated with the SQ test against the other values of δ as well (with the exception of $\delta = 0.10$, which

is enough to push the rejection rate to approximately 100% with both testing approaches).

Figure 4. Power Analysis - SQ Test



Note: Figure 4 plots the rejection frequencies for our 16 models, estimated with the inclusion of Fama-French/Carhart factors. The parameter δ is the level of the event effect. Rejection is based on the SQ test–comparing estimated event-date excess returns for each simulation replicate to the 25^{th} most negative estimated excess return from the estimation period for that replicate.

As with the standard approach, the results for the SQ test indicate that specifications with

lower prediction variance also have greater power for the smaller true event effects. However, the power performance increase across specifications is smaller than–perhaps about half that of–the performance increase we obtain simply by switching to the SQ test. For example, using the SQ test with the market model yields better power against $\delta = 0.01$ than does the best-performing specification (ENR-U) with the standard approach.

5. Further Results

In this section we investigate whether the price of variance reduction appears to be substantial increases in the bias of estimated excess returns. Second, we consider whether it is possible to predict, based on the empirical excess return, for which firm-date pairs ML algorithms will tend to overperform relative to the simple market model.

5.1. The Bias-Variance Tradeoff

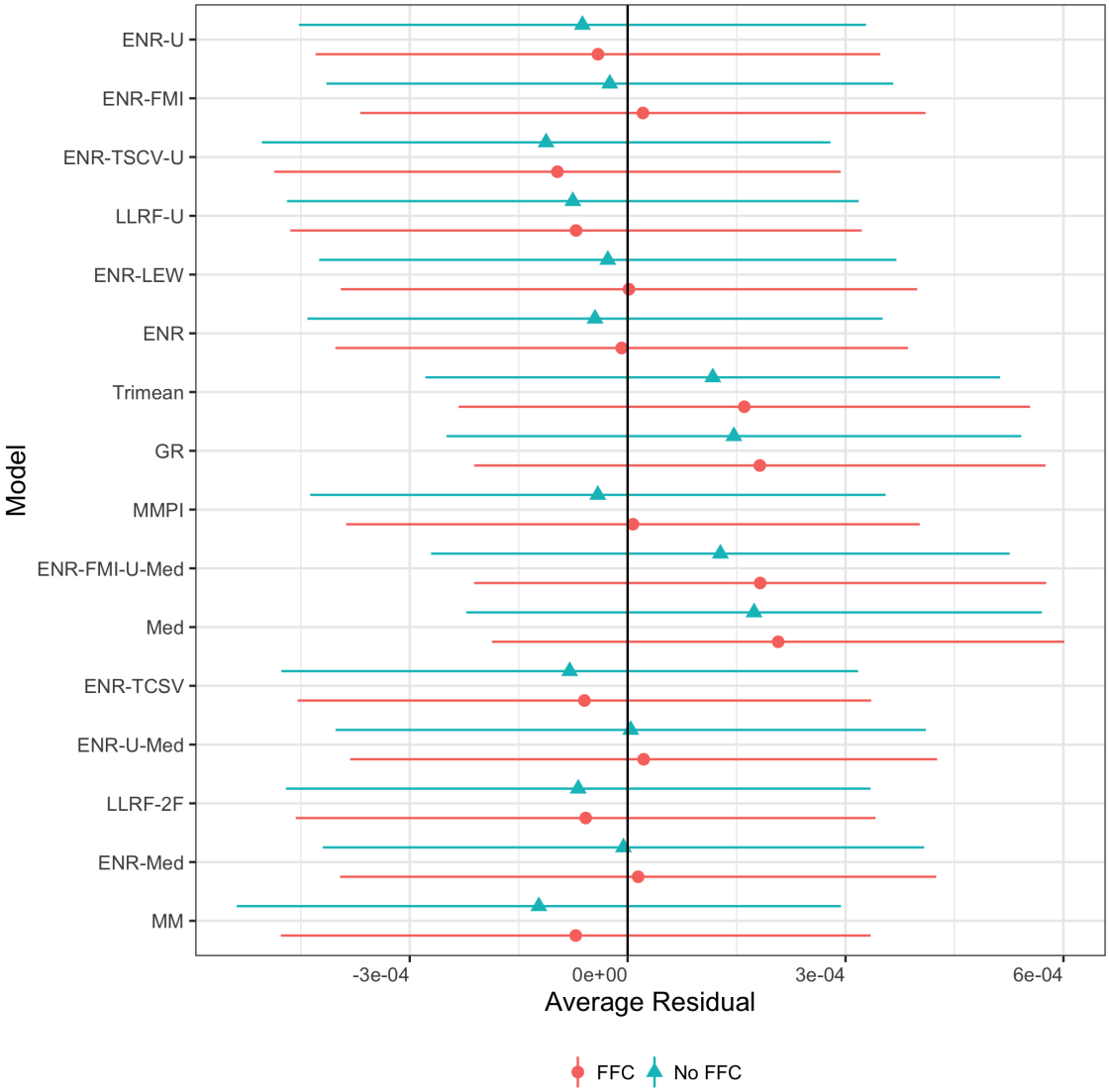
Machine learning models tend to do better at prediction by allowing some in-sample bias in return for reduced variance. As long as the increase in the squared bias is smaller than the reduction in variance, the net impact will be a reduction in mean squared error, because this is the sum of squared bias and variance.

If the induced bias pushed systematically in one direction, that could be a problem in the litigation context, because the disfavored litigant could reasonably argue that ML algorithms were biased specifically against that party. Even if it is desirable to allow additional bias in general in order to reduce variance, fairness in litigation serves as an additional constraint.

On the other hand, bias induced via regularization and other techniques might average out to zero across situations. In that case, ML methods would not predictably and inappropriately hurt one side or the other in litigation. Happily, this is an empirically testable possibility. To test it, we calculated for each specification k the average value of the estimated

event-date excess return from each of the 10,000 simulations we conducted.[28]

Figure 5. Average Residual By Model



Note: Figure 5 plots the average event-date estimated excess return and the 95% confidence level for our 16 models, with and without Fama-French/Carhart factors.

Figure 5 shows these averages. The quantile regression-based specifications seem to have residuals with means above 0, and the other specifications tend to have residuals with means

[28]These averages are not identically 0 for any specification, because they involve out-of-sample estimated residuals rather than in-sample ones. Of course in-sample estimated residuals will have mean exactly equal to 0 for specifications that use least-squares.

below 0. But the magnitudes involved are basically trivial small–even the greatest mean deviation from 0 appears to be no more than 0.0002, i.e., representing an increment to daily returns of just 2 basis points. We conclude that whatever bias is induced by regularization or random forest methods is for practical purposes unimportant.

6. Conclusion

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 firms' securities. Event studies are also widely used in civil litigation, with billions of dollars in settlements ultimately hinging on the outcome of a potentially flawed exercise. It is now well understood that because litigation-relevant studies usually involve only a single date, those conducting event studies for litigation use should modify techniques created for academic use in appropriate ways, especially when those techniques rely importantly on normality assumptions or central limit theorem applicability. It is also understood that single-firm event studies have various problems related to the relatively high excess return variance they involve.

In this paper we explore whether various machine learning and other robust-estimation techniques can be used to enhance the predictive power of excess return calculations in event studies conducted on single securities for securities litigation. We find that estimation with regularization (also called penalization) can yield reductions in event-date excess return variance and improvements in test power. Our best-performing specification reduces event-date excess return variance by about 15% relative to the simple market model with no other variables included. It also has greater power, with improvements in rejection rates on the order of a few percentage points against moderately sized true event effects (e.g., 1-3 log points).

Although these modest gains could be valuable, they are smaller than performance improvements realized by other modifications of the simplest market model. First, simply

including a peer index based on returns for firms in related industries appears to make quite a large difference in prediction variance, and a noticeable one in test performance. Including the Fama-French/Carhart factors also brings improvement, although this is relatively small once a peer index is included.

Second, performance on significance tests is markedly better using the robust SQ test proposed by Gelbach et al. (2013) than when using the standard t-test approach with critical values based on the normal (or Student's t) distribution. Using the SQ test basically eliminates size distortions that plague the standard approach,[29] and it also yields substantial power improvements for smaller true event effect sizes.

In sum, our findings indicate that ML methods can improve single-firm event study performance in ways that could matter in litigation, but they also show that ML methods are less important than previously suggested improvements. Of course there is no reason one couldn't, nor, thus, shouldn't take advantage of both those earlier improvements and ML methods, and that is our advice.

[29]The random forest-based specifications are an exception to this general rule; as noted we think there may be a glitch in the optimization routine we used for these.

References

Baker, A. C., 2016. Single-firm event studies, securities fraud, and financial crisis: problems of inference. Stanford Law Review 68, 151–234.

Binder, J. J., 1998. The Event Study Methodology Since 1969. Review of Quantitative Finance and Accounting 11, 111–137.

Brav, A., Heaton, J., 2015. Event Studies in Securities Litigation: Low Power, Confounding Effects, and Bias. Washington University Law Review 93, 583.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Brown, S. J., Warner, J. B., 1980. Measuring Security Price Performance. Journal of Financial Economics 8, 205–258.

Brown, S. J., Warner, J. B., 1985. USING DAILY STOCK RETURNS The Case of Event Studies. Journal of Financial Economics 14, 3–31.

Campbell, J. Y., Lo, A. W., Mackinlay, C. A., 1997. The Econometrics of Financial Markets.

Carhart, M. M., 1997. On Persistence in Mutual Fund Performance. The Journal of Finance 52, 57–82.

Chandra, R., Moriarity, S., Lee Willinger, G., 1990. A Reexamination of the Power of Alternative Return-Generating Models and the Effect of Accounting for Cross-Sectional Dependencies in Event Studies. Journal of Accounting Research 28, 398–408.

Corrado, C., 1989. A Nonparametric Test for Abnormal Security-Price Performance in Event Studies. Journal of Financial Economics 23, 385–395.

Corrado, C. J., 2011. Event studies: A methodology review. Accounting and Finance 51, 207–234.

Dove, T., Heath, D., Heaton, J. B., 2019. Bias-Corrected Estimation of Price Impact in Securities Litigation. American Law and Economics Review 21, 184–208.

Fama, E. F., Fisher, L., Jensen, M. C., Roll, R., 1969. The Adjustment of Stock Prices to New Information. International Economic Review 10, 1–21.

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

Fisch, J. E., Gelbach, J. B., Klick, J., 2018. The Logic and Limits of Event Studies in Securities Fraud Litigation. Texas Law Review 96, 553–621.

Friedberg, R., Tibshirani, J., Athey, S., Wager, S., 2018. Local Linear Forests pp. 1–36.

Gastwirth, J. L., 1966. On Robust Procedures. Journal of the American Statistical Association 61, 929–948.

Gelbach, J. B., Helland, E., Klick, J., 2013. Valid Inference in Single-Firm, Single-Event Studies. Tech. Rep. 2.

Hein, S. E., Westfall, P., 2004. Improving Tests of Abnormal Returns by Bootstrapping the Multivariate Regression Model with Event Paraments. Journal of Financial Econometrics 2, 451–471.

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction Policy Problems . American Economic Review: Papers & Proceedings 105, 491–495.

Koenker, R., Bassett, G., 1978. Regression Quantiles. Econometrica 46, 33–50.

Kothari, S., Warner, J. B., 2007. ECONOMETRICS OF EVENT STUDIES. In: *Handbook of Empirical Corporate Finance*, vol. 1, pp. 3–36.

Mackinlay, C. A., 1997. Event Studies in Economics and Finance. Journal of Economic

Literature XXXV, 13–39.
Mitchell, M. L., Netter, J. M., 1994. The Role of Financial Economics in Securities Fraud Cases: Applications at the Securities and Exchange Commission. Tech. Rep. 2.
Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Source: Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.
Wooldridge, J. M., 2002. Econometric analysis of cross section and panel data. MIT Press.