Journal of Law, Finance, and Accounting, 2020, 5: 231–272

Machine Learning and Predicted Returns for Event Studies in Securities Litigation

Andrew Baker[1] and Jonah B. Gelbach[2]

[1] *Stanford University, Stanford, CA, USA; abaker2@stanford.edu*
[2] *University of California at Berkeley, Berkeley, CA, USA; gelbach@berkeley.edu*

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 abnormal returns and significance-test power. Thus we believe that 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 *et al.* (2013), because it is robust to the considerable non-normality present in abnormal returns.

Keywords: Event studies/market efficiency studies, financial econometrics, asset pricing

JEL Codes: K220, G120

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

ISSN 2380-5005; DOI 10.1561/108.00000047

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 *et al.*, 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. And even in a market characterized by less than perfect informational efficiency, it is reasonable to expect that newly available material information will affect price quickly. 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 abnormal returns are normally distributed, the Student's t distribution is correct in finite samples. Unfortunately, there is considerable evidence that abnormal returns are not in fact 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 effect—which is an average of a sort—will be approximately normal. However, 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 one 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 *et al.* (2018) explore the consequences of different study design decisions, using the Halliburton case to show how attention to oft-ignored methodological issues can have substantial implications for case determinations.

This literature deals primarily with the inferential properties of single-firm event studies, that is, how significance tests for event-date abnormal 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 abnormal 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 abnormal return.[3] Given that the abnormal 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 abnormal return itself.[4]

A main thrust of our argument in this paper is that event studies, as used in securities litigation, can be viewed as out-of-sample prediction problems. This is important because modern machine learning (ML) methods have proven quite

[1] An exception is Dove *et al.* (2019), who focus on issues involved in damages estimation.

[2] A different question is whether classical statistical significance testing is the right approach to assessing whether there was price impact. Work by Gelbach and Hawkins (2020) addresses this question from a Bayesian perspective. Gelbach and Fisch (2021) take an alternative approach, suggesting that an allowable Type II error rate be specified for benchmark fraud situations, with the Type I error rate determined by the nature of the defendant firm's abnormal returns distribution. For exposition's sake, we ignore these interesting points in this paper, although they could be incorporated in applications.

[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] See Dove *et al.* (2019).

useful for such problems; see, for example, Kleinberg *et al.* (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 three additional Fama-French and Carhart factors, as well as any other variables that might be used by a sufficiently flexible prediction function.[5] We assume that X_{it} includes a 1, which allows for a nonzero intercept if a specification's algorithm would choose that result. 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 though by design some g^j functions will effectively ignore some predictors. In all of the specifications we consider below, the predictor variables enter linearly. Thus, we can write specification $g^j(X_{it}) = X_{it}'\beta^j$ for some coefficient vector β^j.

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

$$r_{it} = X_{it}'\beta^j + \zeta_{it}^j. \tag{1}$$

Given estimated coefficient vectors $\widehat{\beta}^j$, the mean squared error of the predicted abnormal return for specification j, $\widehat{\zeta}_{it}^j = r_{it} - X_{it}'\widehat{\beta}^j = X_{it}'(\beta^j - \widehat{\beta}^j) + \zeta_{it}^j$, may as usual be decomposed into the sum of a squared bias term and a variance term (each conditional on the relevant data; we suppress the conditioning information for notational clarity). Thus the mere fact that $Bias(\widehat{\zeta}^2)^2 > Bias(\widehat{\zeta}^1)^2$ does not tell us the ordering of mean squared error. It is possible that $V(\widehat{\zeta}_{it}^2)$ is enough lower than $V(\widehat{\zeta}_{it}^1)$ to make up for the difference in bias. For example, under certain conditions the standard OLS estimator is known to be the minimum-variance unbiased predictor. But that does not make

[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 French promote 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.

it the minimum MSE predictor, because enforcing unbiasedness is equivalent to imposing a constraint on estimation that requires $Bias^2 = 0$. Relaxing this constraint will allow some additional bias, but with the potential payoff of a large enough reduction in variance to reduce mean squared error. 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 seriously such possibilities by considering the MSE performance of a large variety of return models.

We note that MSE performance can be improved in multiple ways. One is to provide a better functional form of the predicted return given data X_{it}, for example, by either freeing up or constraining the value a regressor's coefficient is allowed to take. For example, some specifications we consider restrict the coefficients so that peer firms' returns enter only through an equally-weighted index, whereas others allow each potential peer firm's return to have its own coefficient value. Another way to improve MSE performance is to use a better way to estimate the parameters of a given model. Familiar examples outside the machine learning literature are linear regression models with non-sphericality, which might be estimated using FGLS to reduce the variance of the coefficient estimator, and quantile regression-based estimation when there is non-normality in residuals. The approaches we use in this paper fall into both categories. Although all specifications we consider involve linear mappings from the set of potential predictors to the target firm's stock returns, some of these specifications involve constraints that others don't impose. The specifications also differ in the regularization and optimization methods used. To avoid conflating the distinct concepts of parameter constraints and and estimation method, we use the word "specification" to refer to the combination of both types of choices.

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 for prediction purposes (Tibshirani, 1996). As noted above, OLS provides the best unbiased linear prediction, but it does so at the price of greater out-of-sample variance compared to other methods. That can lead to comparatively poor prediction accuracy in the MSE metric. 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. We find that specifications using penalized regression generally perform well. 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 is 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 conducted any formal tests, a summary measure is the relative out-of-sample MSE of predicted abnormal 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 on this metric makes use of both penalized regression and data-driven peer firm choice. Its out-of-sample variance of predicted abnormal returns is about 87–88% of the out-of-sample variance of predicted abnormal 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. That said, a 12–13% reduction in variance is perhaps best described as modest, rather than huge.

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 abnormal 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

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

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 conventional practice, we first use the standard approach based on normal critical values.[6] Gelbach *et al.* (2013) point out that this approach is invalid if abnormal returns are not truly normally distributed. They show that there are nontrivial consequences of using the standard approach with real-world abnormal returns, whose non-normality is widely known. Accordingly,

[6]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.

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. However, in this paper we use a more recent testing period and restricted set of firms, and we consider a 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.

Second, we find that the specifications that had best performance in the MSE metric also have noticeably better performance in the testing metric. Third, though, this power difference is less than the improvement brought by using the SQ test rather than the standard t-test approach; the latter difference can be substantial for intermediate values of δ.

In sum, we find that ML specifications can improve on standard ones, although to what might best be described as a moderate degree. We also find additional evidence reinforcing the importance of other good practices in conducting event studies for securities litigation. One is the inclusion of some reasonable peer index. A second is 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 standard critical values from the Student's t distribution, improves the performance of ML specifications considerably. A final contribution of this paper is to show that machine learning methods can usefully serve as a basis for choosing which peer firms to include in an event study. As we discuss in Section 3, this reduces what we term "expert degrees of freedom", thereby mitigating the battle-of-the-experts problem, at least partially.

2 Prior Literature

Event study methodology in finance began with a paper by Eugene 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.[7] 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; Brown and Warner, 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 *et al.* (1990) showed that the relative equivalence in performance between the OLS/market model specification and simpler approaches was a statistical artifact of that specification implementation. Moreover, subsequent research verified that abnormal 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, for example, 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.

As discussed in the introduction, 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 abnormal returns from the market model regression and determines

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

whether the event-date abnormal return is more negative than the α-quantile of the empirical distribution of estimated abnormal returns from the pre-event window.[8] 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 abnormal 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 abnormal 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.[9] 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 abnormal return itself. This is the province of the present paper.

3 The Benefits of Data-Driven Methods for Event Studies Used in Litigation

In this section, we argue that the litigation setting is particularly well-suited to the benefits of data-driven methodologies. Under the status quo approach, plaintiffs and defendants hire expert witnesses who proffer empirical evidence to the court regarding the relative merits of each side's position. This inevitably devolves into a "battle of the experts", characterized by a seemingly intractable divide between opposing experts over complicated and often subjective analytical decisions (Haw, 2012). A generalist court–or jury, in the rare cases that get to one–is then required to settle the dispute between competing experts, often leading to erroneous or case-specific decisions being built into precedent.

The data-driven approaches we explore in this paper have the advantage that they eliminate certain subjective choices made by opposing experts in favor of objective determinations of model fit based on a transparent and reasonable

[8] For an upper-tailed version, one determines whether the event-date abnormal return is greater than the $(1-\alpha)$-quantile; for a two-sided version, one determines whether the event-date abnormal return is between the $(\alpha/2)$-quantile and $(1-\alpha/2)$-quantile.

[9] We do not implement the GARCH approach in this paper, but we believe one could do so with appropriate modifications.

metric, out-of-sample prediction error. To be sure, this advantage applies not only in adversarial litigation, but also in scholarly economic research, which in recent years has begun to move toward data-driven decision procedures, for example, with case study methods and the selection of nuisance parameters in settings with many potential control variables (Abadie *et al.*, 2010; Belloni *et al.*, 2013; Chernozhukov *et al.*, 2018; Athey *et al.*, 2018).

Transparent data-driven methods reduce what we call "expert degrees of freedom". In the securities litigation setting, these methods thus reduce the ability of experts to generate differences in testimonial assessments solely on subjective modeling choices or data restrictions. We focus here on the composition of firms that enter the peer index used in event studies.

Experts testifying in securities litigation have often taken a different approach. Many of these experts advocate controlling for both overall market factors, though they do not usually include the FFC factors, and industry-level factors. In these studies, the two-factor model that includes returns for the broader market as well as a set of firms in the firm's own industry has become an empirical standard. Each expert selects a set of firms to include in an industry peer index, and each argues that the resulting index captures industry-wide movements in asset returns that must be accounted for when estimating the counterfactual abnormal performance for the defendant firm on event dates in question. Such arguments may be based on seemingly compelling bases such as cross-firm correlation or similarities in the firms' businesses. But because experts know who hired them, there is a reasonable concern that peer-index choices may have been made in part due to bias, whether conscious or unconscious. Differences in experts' choice of peer selection drive differences in index construction, and hence differences in the calculation of abnormal performance. Judges and juries typically lack the expertise or information necessary to determine which expert has selected the more appropriate peer index. The result is that in the status quo, fact-finding about abnormal returns may be importantly driven by the ability of experts to persuade in ways unrelated to econometric quality.

A data-driven approach can eliminate this problem. Rather than allowing experts to subjectively choose a set of firms to serve as industry peers, data-driven approaches leverage patterns in the underlying data to tell us explicitly which return series are best able to predict the stock return using held-out data. Courts should be interested in the best estimate of the counterfactual return prediction, rather than which peers do or don't enter an expert's index. If it is possible to agree on a performance metric–a metric for determining which peer index composition is "best"–then it will be possible to use data-driven peer index composition. Showing how to do this is an important contribution of this paper.

To demonstrate, assume there are two experts in a case — A and B. Both conduct event studies to determine the abnormal returns for $StockX$ on date d_0. Expert A selects a set of firms, G^A, and a peer index, P^A, calculated as the equally-weighted average of the returns on the firms in G^A; Expert B

selects corresponding firms and peer index G^B and P^B. Each then uses OLS to estimate the relationship between the defendant's abnormal returns and returns of market and peer indices. The experts estimate the parameters of

$$r_t^X = \alpha^k + \beta_1^k \times mkt_t + \beta_2^k \times p_t^k + \zeta_{it}^k \quad k \in \{A, B\},$$

where r_t^X is the return on $StockX$ on day t, α^k is a constant (the expected return on the stock when the market and respective peer index returns are 0), mkt_t is the return on the aggregate market index, and p_t^k is the return series for the peer indices. After conducting this event study, each expert reports the abnormal return on the event date, d_0, as the difference in the realized return and the predicted return from the fitted coefficients from their event study model:

$$\hat{\zeta_{d_0}^k} = r_{d_0}^X - \left[\hat{\alpha}^k - \begin{pmatrix} mkt_{d_0} & p_{d_0}^k \end{pmatrix} \begin{pmatrix} \hat{\beta}_1^k \\ \hat{\beta}_2^k \end{pmatrix}\right].$$

A court or jury is then required to determine whether they find $\hat{\zeta}_{d_0}^A$ or $\hat{\zeta}_{d_0}^B$ more credible in light of the argument made for industry-index construction by each expert.

A data-driven approach can reduce the scope of such discretion by using a pre-specified algorithm to select which firms among a set G^S of potential peers to include; most algorithms also allow the weights on each included peer firm to be set in a data-driven manner. The set of potential peer firms, G^S, could be the union of G^A and G^B, or perhaps all firms within a certain SIC-code industry. What is important is that this set be broad enough to include all firms that might reasonably serve as peers. Then a data-driven method can be applied, allowing *each* of the firms in G^S to enter the ultimate abnormal returns model with whatever weight minimizes the algorithm's objective function. The algorithm does this selection using the estimation-period returns data and applying cross-validation–multiple estimation passes using "folds" of held out data–to determine optimal coefficients for the final least-squares estimation.

We think there are numerous potential benefits from this approach over subjective peer-firm selection by experts:

- Data-driven methods use a transparent objective function which calculates predicted returns based on a clear measure of interest: the ability of the predictors to explain stock returns. Once the universe of potential peer firms is determined, the minimization problem becomes a deterministic function of the covariance structure in the data, thereby reducing expert degrees of freedom.

- Data-driven methods are flexible, in that they assign weights to potential peer firms rather than assuming all firms contributed equally (or proportionally to their market valuation) to a peer index.

- Data-driven estimates allow optimal trade-offs of variance and bias while avoiding overfitting.
- Data-driven estimates allow for a much larger set of potential peer firms than can be used within a standard unpenalized linear framework, even as they reduce expert degrees of freedom as described above.

4 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 abnormal return of the stock over the event window, and (3) testing for statistical significance of the abnormal return.

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 abnormal return under that specification on date t for firm i. If a constant is present in g^k, then $E[\zeta_{it}] = 0$.[10] As an example, the simple market model specification is

$$g^{MM}(X_{it}) = \alpha^{MM} + M_t\beta^{MM},$$

so that the simple market model puts positive weight only the market-level return (which does not vary with firm index, i), and the associated abnormal 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 event-date return that cannot be explained with available variables in X_{it}. One way to understand ML methods is that they use more flexible g functions for the expected return, compared to g^{MM}. Another is to think of them as using certain nonlinear alterations to the objective function used for estimating the parameters of the function g^k for specification k. As noted above, we use the term "specification" to refer to the combination of the function g–and, thus, the ways various predictor variables are allowed to enter–the objective function, and the cross-validation algorithm used.

[10] We do not impose or assume the stronger assumption, $E[\zeta_{it}^{MM}|\text{mkt}_t] = 0$. If this assumption did hold, then in the absence of temporal dependence, 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.

Below we consider a total of 22 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. We define the period so that date 251 is always the event date. The firms and event dates are randomly selected, so that event date actual and abnormal returns are not systematically related to anything about our specification choices. Of the 22 specifications, 11 include the Fama-French and Carhart factors. The other 11, including the simple market model specification described just above, do not. The exact ways we pick the 250 dates vary a bit with specification, as described in the discussion below.

Write $i251(b)$ to indicate the firm (i)-date $(t = 251)$ combination used as the event-date for simulation replicate $b \in \{1, 2, \ldots, 10{,}000\}$. Then the estimated event-date abnormal return for specification k is $\widehat{\zeta}^{k}_{i251(b)}$. We take four approaches to comparing performance across models–two involving mean-squared error of the estimated event-date abnormal return, and two involving the performance of significance tests based on event-date abnormal return. We discuss details of these approaches below, just before we report the corresponding empirical results.

4.1 Specifications Used

As noted, we consider 22 specifications. There are 11 distinct approaches to estimation, and for each of these we consider two specifications: one that includes the three additional Fama-French and Carhart (FFC) variables, and one that does not. Here we describe the 11 distinct specifications.[11]

For reference, Table 1 provides details 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 and the objective function used.

4.1.1 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 specification for the 250-day estimation window is:

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

[11] In a prior version of this paper we considered a broader set of models, including nonlinear specifications such as random forests, and penalized quantile regression. Because these models represent a somewhat substantial deviation from the standard linear regression models used in practice, and in light of the fact that they did not demonstrate substantial over-performance in our simulations, we have dropped them from consideration.

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

Acronym	Number	Explanatory Variables	Objective Function (Q)/Estimation Method
MM	1	BMI*	Standard^{+} OLS
MMPI	2	BMI*, EWPI*	Standard^{+} OLS
ENR	3	BMI*, EWPI*; Model-generated†	Elastic net regularization
ENR-U	4	BMI*, UPI*; Model-generated‡	Elastic net regularization
ENR-FMI	5	BMI*, FMI, EWPI*; Model-generated^{FMI}	Elastic net regularization
ENR-LEW	6	BMI*, LEWPI*; Model-generated$^{\diamond}$	Elastic net regularization
ENR-TSCV	7	BMI*, EWPI*; Model-generated†	Elastic net regularization
ENR-TSCV-U	8	BMI*, UPI*; Model-generated‡	Elastic net regularization
LASSO	9	BMI*, UPI*; Model-generated‡	Lasso regularization
SYNTH	10	BMI*, UPI*; Model-generated‡	Synthetic control
ID	11	BMI*, UPI*, Model-generated‡	Elastic net regularization

Note:

* The following abbreviations are used for variables (also known as factors):

- "BMI" 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.

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

4.1.2 Specification 2—Market Model + Peer Index (MMPI)

In this simple extension to Specification 1 (MM), 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.[12] 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 abnormal 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).

4.1.3 Specification 3—Elastic Net Regularization with 2 Factor Model (ENR)

We now introduce 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. Specification 3 uses a form of penalized regression objective function known as elastic net regularization. This form allows penalty weight on both the sum of squared coefficients and the sum of their absolute value. Assuming there are p coefficients to estimate, the residuals are $\zeta_{it} = r_{it} - X_{it}\beta$, with β being a $p \times 1$ column vector. Write $\zeta_{it}(c)$ for the residual value when c is used in place of β. Then 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

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

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 ex ante, we optimize over it in the estimation, using a grid-search approach. To obtain our elastic net regularization estimates of specification mean squared error, we do the following for each set of 250 pre-event date observations:

1. Split the data into ten random groupings, known in the ML literature as folds. Repeat this random folding ten different times.

2. For each of the values of $a \in \{0, 0.1, 0.2, \ldots, 1\}$ 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 $\lambda^*(a)$. Then compute the average value of the MSE corresponding to $c^*(a)$ and $\lambda^*(a)$ over the ten random sets of folds; denote this $\overline{MSE}(a)$.

3. Denote as a^* the value of a that yields the lowest $\overline{MSE}(a)$ among the values from step 2, average over t.

4. Set $\hat{\beta}^{ENR}$ equal to the estimates of the coefficients c with $a = a^*$ and $\lambda = \lambda^*(a^*)$, that is, $\hat{\beta}^{ENR} = c^*(a^*)$.

5. Calculate the estimated abnormal returns using $\hat{\beta}^{ENR}$, that is, $\widehat{\zeta}^{ENR}_{it(b)} \equiv r_{it} - \hat{\beta}^{ENR} X_{it}$.

Note that because we allow $a \in \{0, 1\}$ in step 2, Specification 3 is strictly more general than either lasso or ridge regression; if it does not select either value of a, that means neither model can be optimal.

4.1.4 Specification 4—Elastic Net Regularization with Unconstrained Peer Firm Returns (ENR-U)

This specification generalizes Specification 3 (ENR) by relaxing the constraint that peer firms' returns enter the specification through an equally weighted returns index. Specification 4 (ENR-U) drops that constraint and estimates a distinct coefficient for each peer firm's daily return. Notice that this specification nests Specification 3, because the peer firm index can be achieved by setting the coefficient on each of the N_{peer} peer firms' returns equal to N^{-1}_{peer}. Thus specification 4 is a more flexible specification in terms of index creation

than any specification whose regressor set is the market return and equally weighted peer index.[13]

For Specification 4, we implement the unconstrained peer firm returns specification using the same elastic net regularization approach as in Specification 3. This means the vector of coefficients c used to calculate estimates of $\zeta_{it}(c)$ has $2 + N_{peer} + k$ dimensions—one for a (the intercept), one for c_1 (the market return), one for each of the N_{peer} firm returns, and $k = 3$ for specifications that include the 3 FFC factors, and 0 otherwise.

4.1.5 Specification 5—Regularization All Peer Firms and Forced Market Inclusion (ENR-FMI)

This specification augments Specification 4 (ENR-U) by forcing the estimation process to include the daily return of the market index variable. Depending on the data, Specification 4's algorithm might drop the market index regressor before calculating final coefficient estimates. Specification 5 (ENR-FMI) differs from specification 4 only by preventing this from occurring, so that 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 abnormal returns. Calculating the ENR-FMI coefficient estimate is done using the same method as in specification 4, but with the penalty terms being $\lambda \cdot v_j \left(\frac{1-a}{2} c'c + a \sum_{j=1}^{p} |c_j| \right)$ where v_j is equal to 0 for the market index and 1 for all other regressors. Notice that ENR-FMI is a restricted version of ENR-U, in the sense that nothing stops ENR-U from settling on the ENR-FMI coefficients.

4.1.6 Specification 6—Two-Factor Model with Lasso-Based Equally Weighted Index (ENR-LEW)

Specification 6 involves three steps. The first step of Specification 6 (ENR-LEW) can be thought of as a version of specification 4's (ENR-U) elastic net regularization, except with a set to 1, so that we do lasso, which yields a set of selected peer firms. In the second step, we take the first-step-selected firms and use them to calculate an equally weighted peer index. The third step is to use OLS estimation with this equally-weighted peer index whose constituent firms were selected via the first-step lasso estimation. Specification 6 differs from

[13] 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.

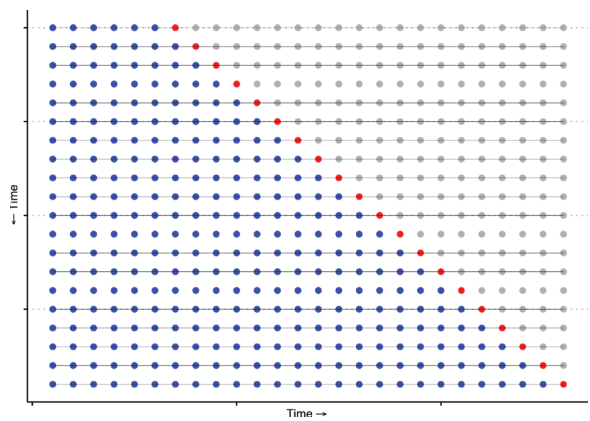
Specification 4 because in the latter specification we use the ENR-estimated coefficients to weight peer firms' daily returns, whereas in Specification 6 we throw out those coefficient estimates and effectively replace them with $1/k$, where k is the number of firms selected for inclusion in the lasso step. We also include the broad market index variable as a final-step predictor in the OLS estimation.

4.1.7 Specification 7—Two-Factor Time-Series Cross-Validation (ENR-TSCV)

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 7 (ENR-TSCV) uses the "evaluation on a rolling forecasting origin" method.

In this procedure, a series of test sets consisting of a single observation are used for cross-validation. The corresponding training set consists of only those observations that occurred *prior* to the observation that forms the test set. 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 7 includes two factors–the broad market index and the equally weighted peer index–and uses the elastic net regularization objective function. Its only difference from Specification 3 is that Specification 7 uses the time-series cross-validation method described above, whereas Specification 3 (ENR) uses ordinary cross-validation.



4.1.8 Specification 8—Time-Series Cross-Validation with Market Index and All Peer (ENR-TSCV-U)

Specification 8 (ENR-TSCV-U) is the same specification as Specification 7 (ENR-TSCV), 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 4 (ENR-U) but with time series cross-validation.

4.1.9 Specification 9—Lasso Regularization with Unconstrained Peer Firm Returns (LASSO)

Specification 9 (LASSO) is the same as Specification 4 (ENR-U), except that instead of using the grid search to implement elastic net regularization, we simply set a to 1. We then use the imputed weights from the unconstrained lasso model to predict returns out of sample.[14]

4.1.10 Specification 10—Synthetic Control (SYNTH)

Specification 10 (SYNTH) uses the synthetic controls approach from Abadie *et al.* (2010), which uses "data-driven procedures to construct suitable comparison groups" that "reduce discretion in the choice of the comparison control units, forcing researchers to demonstrate the affinities between the affected and unaffected units using observed quantifiable characteristics" (pp. 493–494). Stock return event studies rely solely on values of the control-unit variables, that is, there are no control-specific covariates included as predictors. Thus, the synthetic control method can be simplified to a "constrained regression" approach as explained in Imbens and Doudchenko (2016).

The constrained-regression SYNTH specification also chooses weights for the vector of included control variables, r_{it}^C, used in predicting the stock return of the target firm, r_{it} . The control variables are the broad market index return, the returns of each potential peer firm, and possibly the Fama-French/Carhart factors, using the following optimization routine:

$$\hat{w}^{SYNTH} = \arg\min_{\mu,w} \left\{ \left(r_{it} - \mu - w' r_{it}^C \right)' \left(r_{it} - \mu - w' r_{it}^C \right) \right\}$$
$$\text{s.t. } \mu = 0, \sum_{i=1}^{N} w_i = 1, \text{ and } w_i \geq 0, i = 1, \ldots, N$$

We include μ in the objective function in the previous display even though it is constrained to be zero, in order to facilitate a comparison of the SYNTH

[14]The difference between Specification 8 and Specification 6 is that in Specification 8, we use the lasso-selected coefficients themselves, rather than replacing them with equal weights as we do in the second step of Specification 6.

specification with the ID specification discussed next. The SYNTH specification finds the combination of controls within the convex hull that minimizes the prediction error during the estimation period. It is similar to Specification 2 (MMPI), except that it allows weights on peer firms to vary rather than constraining them to be equal. In order for SYNTH to isolate a unique set of weights (the $\widehat{w}^{SYNTH}$), it must constrain the weights to be positive and sum to 1; it also does not allow an intercept term. Notice that this specification is similar to ENR-U (Specification 4), except that ENR-U allows an intercept and does not constrain the coefficients to be proper weights.

4.1.11 Specification 11—Imbens-Doudchenko (ID)

Specification 11, ID, is an estimator proposed by Imbens and Doudchenko (2016), that allows researchers to relax some of the restrictions in Specification 10 (SYNTH): ID allows the weights to be negative, does not restrict the sum of the weights, and allows for a nonzero intercept. The weights still minimize the distance between treated and control units in the estimation period, using elastic net regularization to deal with a potentially large number of control units.

In fact, the Specification 11/ID objective function is identical to that used in Specification 4/ENR-U; see equation (6). The difference between the two specifications lies entirely in the method used to choose the regularization parameters α and λ. Whereas ENR-U uses standard cross-validation, ID treats each control unit—here, these are peer firm returns and any other included variables—as a pseudo-treated unit to determine the optimal value for the parameters, and then uses those weights to predict the outcome for the pseudo-treated unit in the held out period. The performance of the model is then evaluated by computing the average mean squared error over all predictor variables, and the final values of the tuning parameters are those that minimize this mean squared error.

5 Simulation Results

To test the relative predictive accuracy of the eleven specifications described above, we select 10,000 unique firm-events at random over the period from 2009 to 2019 in the CRSP dataset. As discussed above, we index simulation replications with b, and we denote the randomly selected event date for the b^{th} replication firm with date 251. 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). In addition, to avoid the excessive volatility associated with low-share-price firms, we restrict our sample to observations with a trading price above $5. When selecting random event

dates, we require the security in question to have a complete return series for the 250 trading dates immediately preceding the event date for each simulation replication. As mentioned above, in selecting peers, we use only other firms with complete return series over the same period in the same four-digit SIC industry. If there are fewer than eight such firms, we use peers in the same three-digit SIC industry.

To evaluate specifications' mean-squared error performance, we consider two distinct metrics, and we calculate each over two time periods, with and without the FFC factors. Thus we calculate 8 mean-squared error performance measures for each specification. We describe the details of these simulations below, but first it will be useful to present a summary graph in Figure 1. Each gray circle plots the rank of a specification using one of the performance metrics (these ranks are jittered for readability). The diamonds plot the average rank for each specification over the 8 MSE performance metrics. Blue diamonds represent specifications that use some combination of cross-validated penalized regression and individual peer firms; red diamonds represent specifications that do not use those features. Figure 1 shows that unconstrained penalized regression models generally perform best. Notably, the standard CAPM market model clearly performs the worst of our 11 specifications, even when we include the FFC factors.

We turn now to our detailed results.

5.1 Comparison Approach 1: Abnormal Return Variance Normalized Against the Simple Market Model's Average Event-Date Variance

Recall that for simulation replication b, the firm-event date combination is denoted by $i251(b)$; we denote a generic date for the same firm, $it(b)$. We define the squared value of the abnormal return for specification k on firm-date $it(b)$ as $\widehat{w}^k_{it(b)} \equiv \left(\widehat{\zeta}^k_{it(b)}\right)^2$. Using the convention that the event date is labeled $t = 251$, the out-of-sample prediction of interest for specification k on simulation replicate b is $\widehat{w}^k_{i251(b)}$. The estimated out-of-sample mean squared error (MSE) for specification k over the $b \in \{1, 2, \ldots, 10{,}000\}$ replications is thus:

$$\widehat{MSE}^k_{oos} \equiv \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} \widehat{w}^k_{i251(b)}. \tag{7}$$

For each specification k, we then compute the ratio of of the out-of-sample MSE $\widehat{R}^k_{oos}$ as:

$$\widehat{R}^k_{oos} \equiv \frac{\widehat{MSE}^k_{oos}}{\widehat{MSE}^{MM}_{oos}}, \tag{8}$$

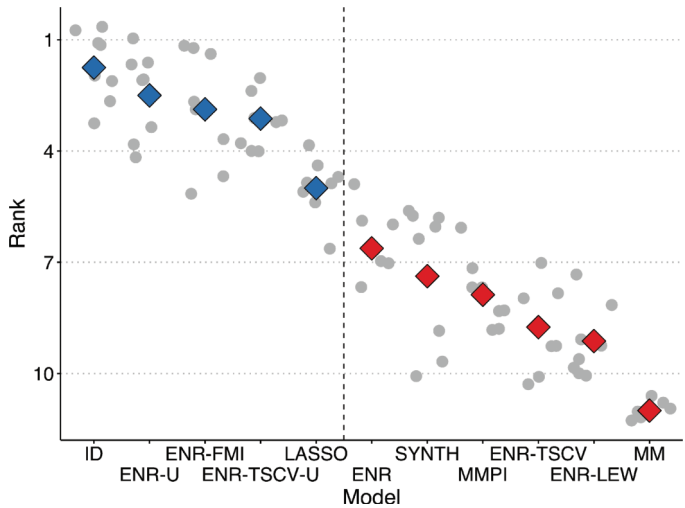


Figure 1: Distribution of Specification Ranks Across Models and Tests.

Note: Figure 1 plots specification ranks. Each specification has 8 MSE performance values: two time periods (1999–2009 vs. 2009–2019), with and without the FFC factors, and two MSE normalization approaches ($\widehat{R}^k_{oos}$ and $\widehat{R}^k_{het}$, described below). Each gray dot represents a rank from 1 to 11, and each rank is represented once for each of the eight time-period/FFC-factor/MSE-metric combinations. The diamonds plot the specifications' average ranks. Blue diamonds signify models that allow firms to enter the regression function individually and use cross-validation and penalized regression; red diamonds represent specifications that do not.

where the denominator is the out-of-sample MSE for Specification 1 without the FFC factors. By construction the ratio in 8 is 1 for that specification, so 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 simple market model without the FFC factors.[15]

Figure 2 plots the values of $\widehat{R}^k_{oos}$ for the 11 models with (triangles) and without (circles) the Fama-French Carhart (FFC) factors. The order in which the specifications are listed on the vertical axis is determined by performance in the specifications with the FFC factors, so that the specification reported in the top row has the lowest value of $\widehat{R}^k_{oos}$, and the one that appears in the bottom row has the highest value. Figure 2 shows that the worst-performing specification without the FFC factors is the simple market model. The best-performing specification, both with and without the FFC factors, is ENR-U (Specification 4). 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 one of the more flexible penalized regression specifications

[15] As we will see momentarily, no specifications has $\widehat{R}^k_{oos} > 1$.

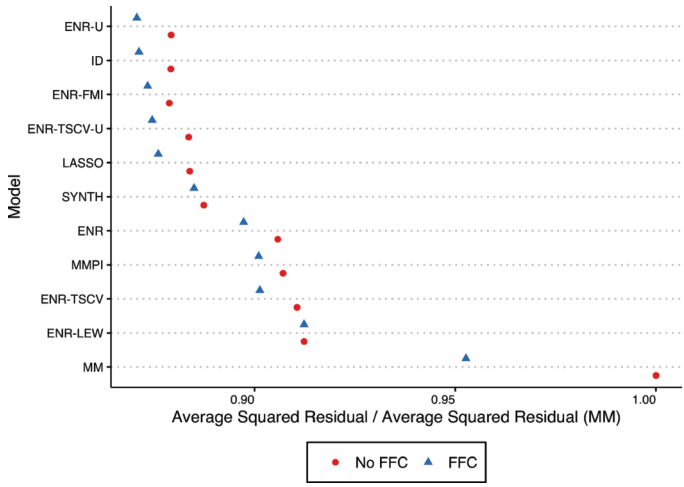


Figure 2: Ratios of Average Squared Residual to MM.

Note: Figure 2 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 FFC models.

we considered; indeed, it nests all the others, so it is not surprising that it performs best.

Other specifications are very close to ENR-U in performance terms, including ID, ENR-FMI, ENR-TSCV-U, and LASSO. The ID model is very close to ENR-U, differing only in its tuning procedure for penalization parameters. The ENR-FMI specification differs from ENR-U only in that it forces the inclusion of the broad market index in the final estimation. The ENR-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. Finally, LASSO is equivalent to ENR-U with the stipulation that the hyperparameter $a = 1$ in the maximization problem. The fact that these specifications perform similarly, and better than the others, suggests the importance of flexibility in the way peer firms enter, and in the use of cross-validation to select parameter values.

Most notably, all of the top performing models allow the peer firms to enter the maximization problem individually, rather than being combined into an index before estimation. When the FFC factors are included, the best performing models have average event-date abnormal 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. Finally, we note that including the FFC factors helps, although including some measure for peer performance seems even more important.

5.2 Comparison Approach 2: Abnormal Return Variance Normalized Against Within-Date In-Sample Variance of the MM Specification

Our second comparison approach is meant to deal with heteroskedasticity in abnormal returns. There are some "event" dates, that is, $i251(b)$ dates, on which important events really did occur, and for unmodeled reasons. Abnormal returns will be especially large on such days.[16] 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 simulation's in-sample MSE for the simple market model. One approach to such a normalization would be to compute the average, across the 10,000 simulation replicates, of the ratio of $\widehat{w}^{k}_{i251(b)}$ to $\widehat{w}^{MM}_{i251(b)}$, that is, the ratio for the simulated event date. That should account for both in-sample (the 250 non-"event" dates $t \in \{1, 2, \ldots, 250\}$) and out-of-sample (the randomly selected "event" date, $t = 251$) heteroskedasticity. But it also runs a risk. On any simulated event date when Specification 1 happens to predict almost perfectly, 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. If that happens for purely random reasons, then our normalized performance metric might be dominated by noise rather than signal. To avoid this kind of effect, we instead use the following metric for each model k:[17]

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

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

Notice that $\widehat{R}^{k}_{oos}$ and $\widehat{R}^{k}_{het}$ differ. The $\widehat{R}^{k}_{het}$ metric normalizes the squared event date abnormal 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). $\widehat{R}^{k}_{het}$ is the average over simulation replications of a ratio, whereas its counterpart $\widehat{R}^{k}_{oos}$ is a ratio of averages.[18] The $\widehat{R}^{k}_{oos}$

[16] Similarly, some estimation windows in our simulations 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.

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

[18] To put it slightly differently, $\widehat{R}^{k}_{het}$ normalizes within dates and then averages, whereas $\widehat{R}^{k}_{oos}$ instead averages across dates and only then normalizes. This is why $\widehat{R}^{k}_{oos}$ is not simply $\widehat{R}^{k}_{het}/\widehat{R}^{MM}_{het}$.

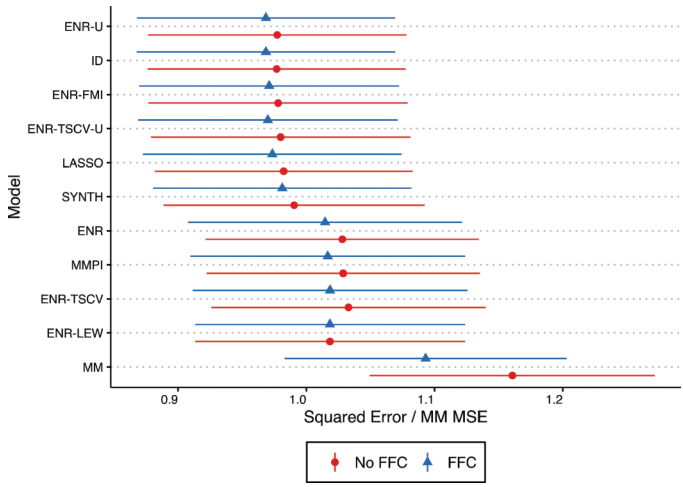


Figure 3: Mean Squared Error By Specification: 10,000 Simulations.

Note: Figure 3 plots the average normalized squared prediction error for our 11 candidate specifications, i.e., $\widehat{R}^k_{het}$. We plot the estimates both with (FFC) and without (No FFC) Fama-French/Carhart Factors.

approach would be problematic if there is so much heteroskedasticity in event-date abnormal 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. Although we believe no constraint forces the specifications to perform similarly with $\widehat{R}^k_{oos}$ as with $\widehat{R}^k_{het}$, Figure 3 indicates that, broadly considered, they do. Because the $\widehat{R}^k_{het}$ denominator is constructed by averaging over a large sample of dates, it does not engender the signal-to-noise problem described above.

Figure 3 plots the mean of the normalized prediction errors of the 11 models, together with 95% confidence intervals. The order in which the specifications are listed on the vertical axis is the same as Figure 2, to facilitate comparisons with the $\widehat{R}^k_{oos}$ results.

Figure 3 shows that the worst-performing specification without the FFC factors is still the simple market model. Notice that its value of $\widehat{R}^k_{het}$ is roughly 1.16, meaning that the out-of-sample variance of predicted event-date abnormal returns for the simple market model is about 16% greater than the in-sample variance for that same model. This difference between in-sample and out-of-sample variance illustrates the empirical importance of overfitting. In addition, the ranking of models follows a similar pattern as in Figure 2: those models that performed best in terms of $\widehat{R}^k_{oos}$ also perform best in terms of $\widehat{R}^k_{het}$. The ENR-U specification again performs best, both with and without the FFC factors; its out-of-sample event-date abnormal return variance is

roughly 97-98% of the in-sample variance of the simple market model; this is good enough for an 84% reduction relative to the 1.16 value for Specification 1 without the FFC factors.

A group of models (ENR, MMPI, ENR-TSCV, and ENR-LEW) are ranked below the top performing in relatively indistinguishably ways. All have out-of-sample average normalized event-date abnormal return variance between 100% and 104% of the in-sample variance of the simple market model, with or without the FFC factors included.

We can draw a number of lessons from Figures 2 and 3. One is that regularization-based ML algorithms appear to reduce event-date abnormal return variance. This conclusion should be tempered in two ways. First is that the 95% confidence intervals in Figure 3 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 highly positively correlated across specifications, so it is difficult to draw meaningful inferences based only on model-specific confidence intervals.

A second lesson from Figure 3 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.

Third, the best performing specifications are either 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). Fourth, 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 specification is not; it performs about the same as other estimators that use either standard cross-validation or don't use regularization at all–e.g., the MMPI specification, which uses unpenalized estimation algorithms but includes both the market model and the equally-weighted peer index. In sum, the best performance in Figures 2 and 3 comes when regularization is paired with some sort of peer-firm adjustment and the FFC factors.

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

We have seen that regularization can enhance the precision of event study abnormal return estimates. We now assess how important precision improve-

ments are for significance tests of whether abnormal returns are different from 0. These tests are important in securities litigation, because class certification and the resolution 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.[19] 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 abnormal return to its estimated standard error is less than the α-quantile of the standard normal distribution:

$$\widehat{T}_b^k \equiv \frac{\widehat{\zeta}_{i251(b)}^k}{RMSE_b^k} < z_\alpha, \tag{10}$$

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 event-date abnormal 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 adjust $\widehat{T}_b^k$ to account for the "true" magnitude of the event effect that is of interest. Suppose events of interest cause firm value to fall by the amount δ; then event-date returns would be $r_{i251}^\delta = r_{it} - \delta$. 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 abnormal 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 10. 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, i.e., for power assessment we can think of the rejection rule as $\widehat{T}_b^k < z_\alpha + \delta/RMSE_b^k$. Because $\delta \geq 0$, it follows that we will reject more frequently the greater is the true effect magnitude, as usual. Finally, we note that size may be thought of the rejection rate when $\delta = 0$.

[19] In our simulations we do not adjust for the degrees of freedom of the model. This is functionally irrelevant for the models that use solely indexes as independent variables, given the number of degrees of freedom we have with 250 estimation dates. However, for models that allow peer firms to enter the optimization individually (e.g. ENR-U or ID), this may not always be the case depending on the number of potential peer firms and the optimized penalty weights. However, addressing this would only lower the actual test size, which Figure 4 shows is already below the desired level. Because degrees-of-freedom adjustments involve a positive monotonic transformation, they do not affect the SQ test's performance, which suggests that analysts should use the SQ test when testing significance using penalized regression based models.

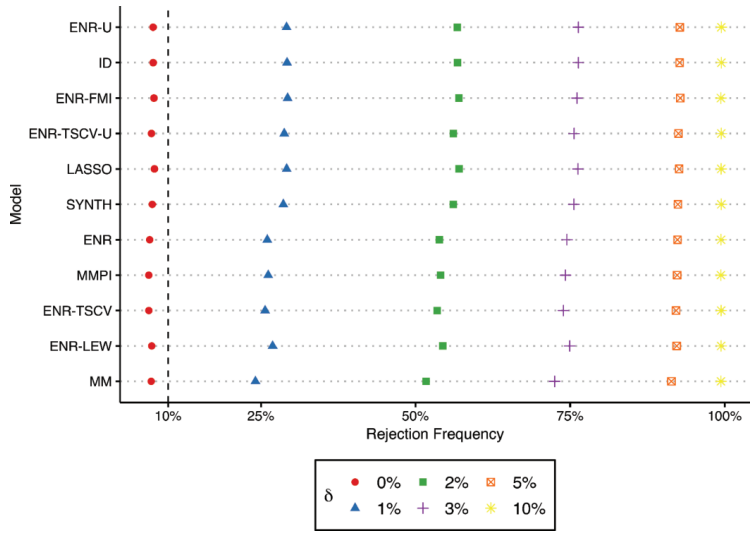


Figure 4: Power Analysis - Standard Approach Tests of Significance.

Note: Figure 4 plots the rejection frequencies for our 11 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.

Figure 4 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 11 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. These findings as to substantial size distortions echo those in Gelbach *et al.* (2013).

For values of δ above zero–i.e., when there really was an event effect–there is some variation in performance across specifications. Most notably 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 2, so this rightward lean indicates that specifications with lower abnormal return variance according to the $\widehat{R}^k_{oos}$ 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. Finally, we note that with effect sizes as large as 5% or 10%, it doesn't much matter which specification one uses–all have substantial power.

5.4 Comparison Approach 4: Significance Test Performance Using the Sample Quantile Test

The poor size of the standard approach tests exhibited in Figure 4 is unsurprising given (i) the well known non-normality of abnormal returns and (ii) the arguments made and evidence provided in Gelbach *et al.* (2013). That paper shows that when abnormal 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 4.

Gelbach *et al.* (2013) propose an alternative, based on the sample quantiles of the empirical distribution function (EDF) of estimated abnormal 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.[20]

Although Gelbach *et al.* (2013) 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 abnormal 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. This location difference is zero under the null hypothesis anyway. The Glivenko-Cantelli theorem then implies that the EDF of estimation-window abnormal returns is a consistent estimator for the true distribution function of the event-date abnormal return. Accordingly, the sample quantiles of the estimation period are consistent estimators for the true quantiles of the event-date abnormal 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 5 reports SQ test rejection rates for the same values of δ investigated using the standard approach tests reported in Figure 4. 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 abnormal return on each simulation replicate to a critical value that equals the 25^{th} most negative estimated abnormal return, because that value is the sample 0.10-quantile of abnormal returns.

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

[20] See also Conley and Taber, 2011, who prove a similar result in a more general setting.

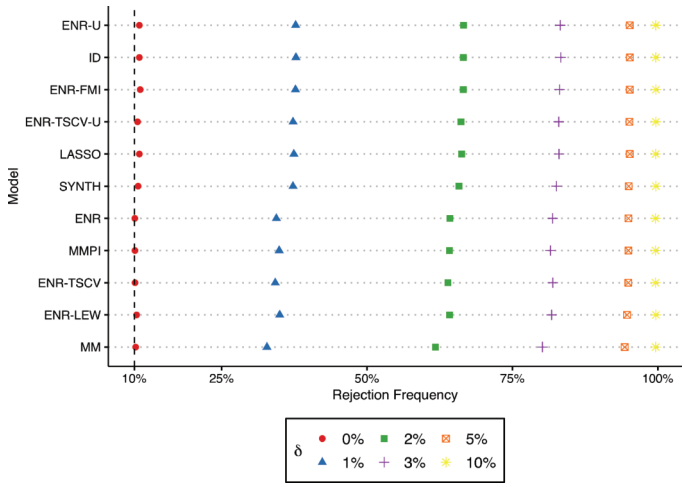


Figure 5: Power Analysis - SQ Test.

Note: Figure 5 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 abnormal returns for each simulation replicate to the 25^{th} most negative estimated abnormal return from the estimation period for that replicate.

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).

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 the performance increase we obtain simply by switching to the SQ test. And note that using the SQ test with the market model (bottom row of Figure 5) yields better power against $\delta = 0.01$ than does the best-performing specification (ENR-U) with the standard approach (top row of Figure 4).

6 Further Results

In this section we investigate whether the price of variance reduction is a substantial increase in the bias of estimated abnormal returns. Second, we test

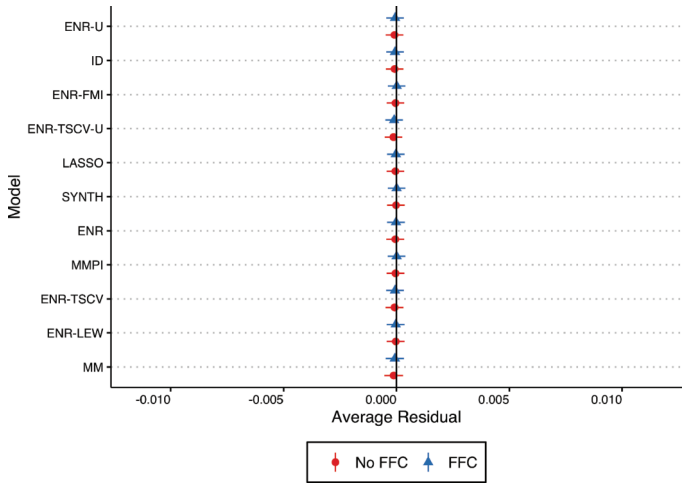


Figure 6: Average Residual By Model.

Note: Figure 6 plots the average event-date estimated abnormal return and the 95% confidence level for our 11 models, with and without Fama-French/Carhart factors. For perspective, the range of the x-axis is set equal to one standard deviation of the return series in our sample.

whether these results hold during a different time period when the volatility of stock returns was noticeably different (1999-2009). Finally, we explore whether regularization-based event study methods can provide performance increases in academic cross-sectional regressions.

6.1 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 so, this could present a problem in the litigation context, because disfavored litigants could reasonably argue that ML algorithms were biased against them. Fairness to the parties is, after all, an additional constraint 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 abnormal return from each of the 10,000 simulations we conducted.[21]

[21] 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 any specification whose final step uses least-squares.

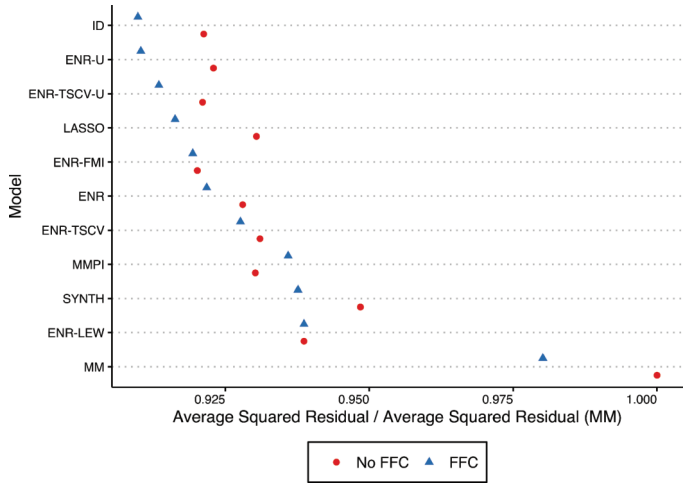


Figure 7: Ratios of Average Squared Residual to MM in Crisis Period.

Note: Figure 7 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) during the Crisis Period of 1999-2009. The models are reported in order of their predictive power in the FFC models.

Figure 6 shows these averages, with the range of the x-axis set equal to one standard deviation of the return series in our sample.. The magnitudes of deviation from zero involved are trivially 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, or roughly one percent of the standard deviation of daily returns. We conclude that whatever bias is induced by regularization is for practical purposes unimportant.

6.2 Results During The Financial Crisis Period

The simulation results in the earlier portion of this paper were conducted during the most recent ten-year period of market return data (2009-2019). While it makes sense to test our prediction models on recent return data, this was a period of comparative market tranquility. In Figure 7 we replicate the analysis from Figure 2 using data from the preceding ten year period (1999-2009), which includes both the dot-com bubble and collapse and the financial crisis. Figure 7 plots the ratio of the average out of sample mean squared error for each model as a percentage of the out of sample mean squared error for the Market Model (MM) specifications without FFC factors ($\widehat{R}^k_{oos}$), and is sorted vertically by model according to the value of $\widehat{R}^k_{oos}$.

The results in Figure 7 are broadly consistent with those in Figure 2. As there, the best performing models here are those that rely on penalized

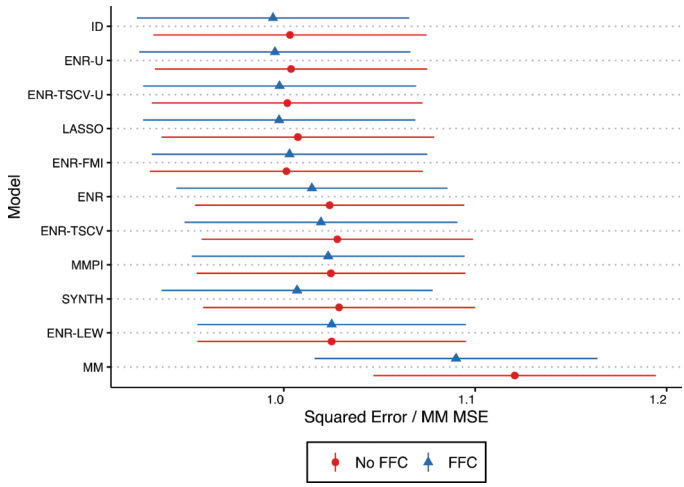


Figure 8: Mean Squared Error By Specification in Crisis Period: 10,000 Simulations.

Note: Figure 8 plots the average normalized squared prediction error for our 11 candidate specifications, i.e., $\widehat{R}^k_{het}$, during the Crisis Period. We plot the estimates both with (FFC) and without (No FFC) Fama-French/Carhart Factors.

regression and which flexibly allow individual peer firms to enter the optimization problem (ID, ENR-U, ENR-TSCV-U, and LASSO). One difference during the financial crisis period is that the constrained-regression synthetic control approach (SYNTH) performs materially worse; indeed it is one of the worst performing models in relative terms. This suggests that relaxing the constraints that require covariate coefficients to be proper weights may be especially worthwhile during periods of market volatility. Lasso also does poorly without the FFC factors, though it performs relatively well with them included.

In Figure 8 we similarly replicate the analysis from Figure 3 for the Crisis Period. Here, we plot estimates of $\widehat{R}^k_{het}$, which normalizes the squared event date abnormal 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). Figure 8 plots the mean of the normalized prediction errors of the 11 models over the Crisis Period (1999–2009), together with 95% confidence intervals. The order in which the specifications are listed on the vertical axis is now the same as Figure 7.

Again we see a general consistency between comparison approaches; the models that perform best in terms of $\widehat{R}^k_{oos}$ also generally perform well in terms of $\widehat{R}^k_{het}$. The best performing models in terms of $\widehat{R}^k_{het}$ use regularization and allow peer firms to enter the objective function individually, although SYNTH performs better in terms of $\widehat{R}^k_{het}$ than $\widehat{R}^k_{oos}$ (especially when including FFC

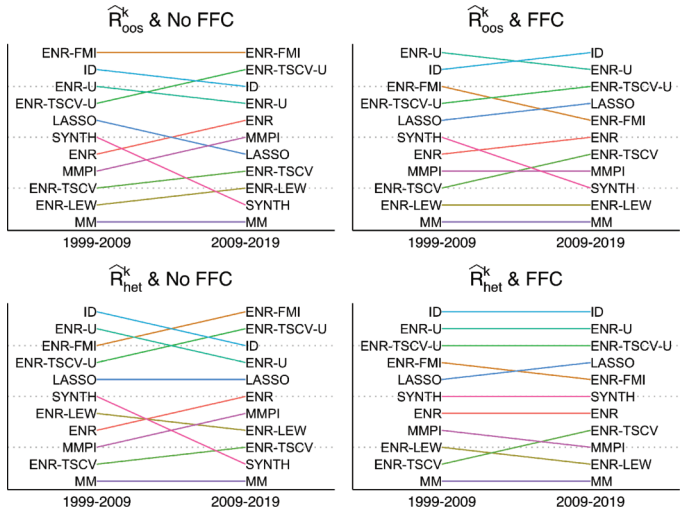


Figure 9: Relative Rankings of Models Across Time Periods.

Note: Figure 9 lists the 11 specifications in order of rank in the two time periods 1999-2009 and 2009-2019. The panels on the left show the relationship between ranks and time period for models without FFC factors using the $\widehat{R}^k_{oos}$ and $\widehat{R}^k_{het}$ comparison measures respectively, while the right panels show the same change in rankings for models using the FFC factors.

factors). In conclusion, it appears there is a consistent relationship between model performance, even over two non-overlapping time periods.

To further demonstrate this correspondence, we map the relative ranking of estimates between comparison method ($\widehat{R}^k_{oos}$ and $\widehat{R}^k_{het}$) and model inclusion type (with and without FFC factors) over the two periods in Figure 9. While there is some evidence of movement across the middle-to-low performing models, the best performing models show consistency across time period and model/test type.

6.3 Application to Cross-Sectional Event Studies

In the cross-sectional setting, analysts can generally rely on averaging over many securities' returns to remove non-normality and reduce variance. However, when the timing of events are concentrated in a short period, known in the literature as event-date clustering, reliance on large sample properties may be insufficient to remove other potential unobserved explanations for changes in stock prices. Kolari and Pynnönen (2010) write, "it is advantageous to use abnormal-return definitions that reduce cross-correlation to a minimum to maximize the power of the test statistics." Because the methods proposed in this paper have the potential to remove more individual-level unexplained variation in asset returns, it is ex ante plausible that their use

could increase the power of cross-sectional event studies to detect abnormal performance.

To test this hypothesis we perform a simulation design similar to those used in the academic event study methodology literature (Brown and Warner, 1980; Brown and Warner, 1985; Kolari and Pynnönen, 2010). We create 1,000 independently drawn portfolios of 50 securities, with a common randomly selected event date. Following Kolari and Pynnönen (2010) we set the "event date" to 250 and use an estimation window of dates 1 through 239. Each security in the sample must have at least 50 returns in estimation period to be included in a portfolio, and cannot have any missing returns over the period -10 to 10 (Kolari and Pynnönen, 2010). For each of the n = 50 securities in the portfolio we estimate a separate event study over the estimation window, using five candidate event study specifications in this paper (MM, MMPI, ENR-U, LASSO, and ID). We chose these specifications to represent a mix of models used in the standard approach (MM and MMPI), as well as a set of better performing models that rely on penalized regression with minor differences in tuning parameters (ENR-U, LASSO, ID). We then calculate an aggregate portfolio-level test statistic for the significance of the randomly selected pseduo event-date using three test statistics described in Kolari and Pynnönen (2010).

The unadjusted cross-sectional t-statistic (UNADJ) is defined as:

$$UNADJ = \frac{AAR\sqrt{N}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} s_i^2(1+d_t)}}$$

where AAR is the average, non-scaled, abnormal return over the $n = 50$ securities in the portfolio, s_i^2 is the regression model residual variance for security i, and d_t is a sampling error correction component of the form $x_t^{'}(X^{'}X)^{-1}x_t$ resulting from the estimation of the regression parameters in the estimation period, where x_t is the vector of explanatory variable values on the event date t, and X is the matrix of explanatory variable values during the estimation period. This represents the classical cross-sectional test statistic used in most academic event studies. This test statistic is given in equation 18 of Kolari and Pynnönen (2010).

Parametric tests based on scaled abnormal returns (abnormal returns divided by the standard deviation of estimation period residuals) have been found to have superior power in detecting abnormal returns (Patell, 1976; Boehmer *et al.*, 1991). Thus we also consider the scaled abnormal returns (SARs) statistic defined as:

$$A_{it} = \frac{AR_{it}}{s_i\sqrt{1+d_t}}.$$

This statistic is given in equation 5 of Kolari and Pynnönen (2010). SARs are used in the test-statistic proposed in Boehmer *et al.* (1991) (BMP), which

defines the following alternative test-statistic:

$$BMP = \frac{\overline{A}\sqrt{n}}{s}$$

where $\overline{A}$ is the average of the scaled abnormal returns for the $n = 50$ securities in the portfolio and s is the cross-sectional standard deviation of the event-date scaled abnormal returns (i.e. $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(A_i - \overline{A}\right)^2$). This test statistic is given in equation 6 of Kolari and Pynnönen (2010).

Kolari and Pynnönen (2010) shows that with event-date clustering, even low levels of residual cross-correlation between securities in the portfolio can lead to overrejection of a true null hypothesis of zero average abnormal returns. They propose a modified version of the BMP test statistic that takes into account this cross-correlation and leads to more powerful tests. The adjusted BMP test (ADJ-BMP) is defined as:

$$ADJ - BMP = BMP \cdot \sqrt{\frac{1-\overline{r}}{1+(n-1)\overline{r}}}$$

where n is again the number of securities in the portfolio (50) and $\overline{r}$ is the average correlation of the model residuals across the 50 securities in the portfolio during the estimation period. This test statistic is given in equation 11 of Kolari and Pynnönen (2010).

Using our simulated portfolio firms and dates, we calculate the cross-sectional test statistics (UNADJ, BMP, and ADJ-BMP) for each of the MM, MMPI, ENR-U, LASSO, and ID specifications. We then compute empirical rejection rates using a two-tailed test of statistical significance and a 5% significance level.[22] We consider not only a true null hypothesis of zero effect, but also a range of true alternative hypotheses involving non-zero mean effects. In each portfolio the event date log return r_{it} is replaced with $r_{it} + k$ for k taking values in $[-0.02, 0.02]$. The rejection frequency is then calculated by test statistic/specification/k combination as the percentage of portfolios for which the null hypothesis of no abnormal return is rejected. A combination with higher power will have higher rejection frequencies for a given level of imputed abnormal performance k.

Table 2 reports empirical rejection rates against the true null hypothesis, i.e., for the $k = 0$ case. The first column of the table shows that when we use the UNADJ test statistic, all but the MMPI specification over-reject substantially. This remains true when we use the BMP test statistic. Evidently neither of these test statistics appropriately deals with intra-portfolio dependence in

[22] Above we used one-sided tests at a 10% significance level. Here we use the two-sided test and the 5% significance level to facilitate comparison with other work in the non-litigation literature.

Table 2: Rejection Rates Against True Null, (Using 5% Significance Level).

Specification	Test-Statistic		
	UNADJ	BMP	ADJ_BMP
MM	0.083	0.081	0.057
MMPI	0.059	0.059	0.058
ENR_U	0.091	0.097	0.055
LASSO	0.095	0.097	0.056
ID	0.090	0.094	0.059

Note: Each entry shows the rejection frequency against a true null hypothesis of zero event-date effect for our five specifications (MM, MMPI, ENR-U, LASSO, and ID) using the three test statistics derived in Kolari and Pynnönen (2010). UNADJ is the standard unadjusted cross-sectional t-statistic common in the literature, BMP is the SAR-based test statistic from Boehmer *et al.* (1991), and ADJ-BMP is the test statistic proposed by Kolari and Pynnönen (2010) that uses both scaled abnormal returns and an adjustment for the residual cross-correlation among treated securities. The rejection frequencies are based on a two-tailed test of statistical significance at the 5% significance level.

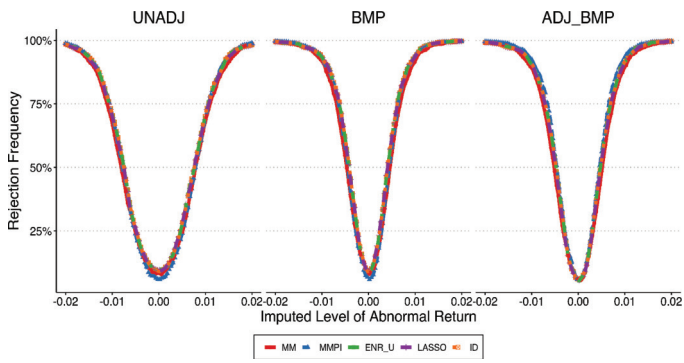


Figure 10: Relative Rankings of Models Over Time Periods.

Note: Figure 10 plots the rejection frequencies for five selected specifications (MM, MMPI, ENR-U, LASSO, and ID) using three test statistics derived in Kolari and Pynnönen (2010). UNADJ is the standard unadjusted cross-sectional t-statistic common in the literature, BMP is the SAR-based test statistic from Boehmer *et al.* (1991), and ADJ-BMP is the test statistic proposed by Kolari and Pynnönen (2010) that uses both scaled abnormal returns and an adjustment for the residual cross-correlation among treated securities. For each test-statistic we calculate the rejection frequencies across our 1,000 portfolios of 50 randomly selected stocks for varying levels of artificially imputed abnormal returns for each specification. The rejection frequencies are based on a two-tailed test of statistical significance at the 5% significance level.

daily returns. By contrast, the ADJ_BMP tests reject between 5% and 6% of the time for all five specifications, well within the 95% confidence interval for the true null hypothesis.[23]

[23]With 1,000 simulation replications, the standard error of the empirical rejection rate under the null hypothesis (that the true rejection probability is 0.05) is $(.05 \times .95/1000)^{1/2} \approx 0.007$, so a 95% confidence interval for the true rejection probability is roughly $[0.036, 0.064]$.

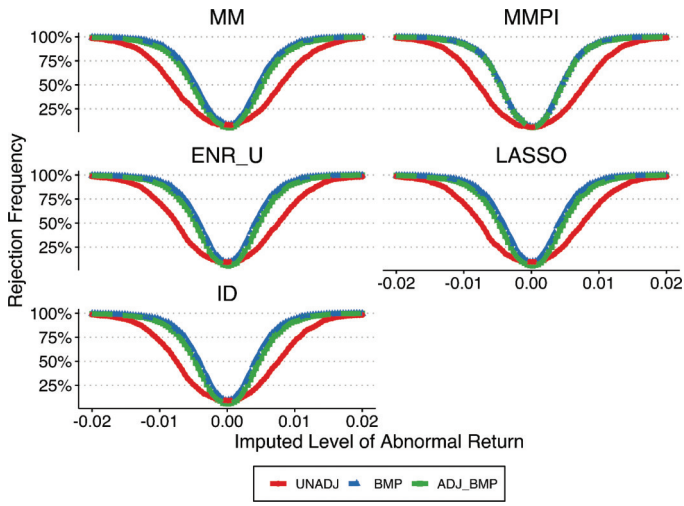


Figure 11: Relative Rankings of Models Over Time Periods.

Note: Figure 11 plots the rejection frequencies for five selected specifications (MM, MMPI, ENR-U, LASSO, and ID) using three test statistics derived in Kolari and Pynnönen (2010). UNADJ is the standard unadjusted cross-sectional t-statistic common in the literature, BMP is the SAR-based test statistic from Boehmer *et al.* (1991), and ADJ-BMP is the test statistic proposed by Kolari and Pynnönen (2010) that uses both scaled abnormal returns and an adjustment for the residual cross-correlation among treated securities. For each of the five specifications we calculate the rejection frequencies across our 1,000 portfolios of 50 randomly selected stocks for varying levels of artificially imputed abnormal returns for each test-statistic. The rejection frequencies are based on a two-tailed test of statistical significance and a 95% confidence level.

These results indicate that there is enough dependence that only the ADJ-BMP test statistic yields reliable inference, regardless of the specification we use (with the exception that for some reason the MMPI specification does well with all three test statistics). Figure 10 plots empirical rejection rates for the five specifications and three test statistics as we vary the true event effect k. For a specification with higher power to detect abnormal performance, the curve in the figure will lie above the curve for other specifications. The figure suggests there are limited power differences across the five specifications, given the test statistic used.

Figure 11, by contrast, depicts substantial power differences across test statistics. Power is considerably lower for the UNADJ test statistic, regardless of specification. The BMP and ADJ-BMP tests have generally similar power, although there is some variation across specifications in their relative performance.

To sum up the results of our cross-sectional portfolio simulation, it is clear that the test statistic matters, but given that one uses the ADJ_BMP test statistic, we find little difference across specifications in either the size or power performance. Thus, our evidence provides little reason to think the

regularization-based methods proposed in this paper for the single-firm event study context would yield much if any performance improvement for portfolio-based event studies, provided that intra-portfolio dependence is accounted for.[24]

7 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 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 abnormal 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 abnormal return calculations in event studies conducted on single securities for securities litigation. We find that estimation with regularization (also called penalization) can yield modest reductions in event-date abnormal return variance and improvements in test power. Our best-performing specification reduces event-date abnormal 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.

[24] We think this result likely can be explained with two observations. First, the daily return variance for a 50-firm portfolio can be expected to be much less than that for a single firm, even if there is substantial (imperfect) dependence. A lower baseline daily return variance might be expected to afford less room for improvement via better peer choice, so that ML methods have less room to improve on out-of-sample performance. Second, for test rejection rates, central limit theory implies that the averaging involved in constructing multiple-firm portfolios will push the daily return distribution toward normality. Thus Student's t critical values naturally should perform better in the portfolio case than in the single-firm case, provided that the standard error is appropriately estimated.

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 eliminates size distortions that plague the standard approach, and it also yields substantial power improvements for smaller true event effect sizes.

In sum, empirical 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. Finally, we emphasize that in light of the moderate improvements presented above, arguably the most important argument in favor using ML methods is their ability to reduce expert witnesses' degrees of freedom by providing an objective basis for determining how to control for peer-firm performance.

References

Abadie, A., A. Diamond, Hainmueller, and Jens. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program". *Journal of the American Statistical Association.* 105(490): 493–505.

Athey, S., G. W. Imbens, and S. Wager. 2018. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions". *Journal of the Royal Statistical Society. Series B: Statistical Methodology.* 80(4): 597–623.

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

Belloni, A., V. Chernozhukov, and C. Hansen. 2013. "Inference on treatment effects after selection among high-dimensional controls". *Review of Economic Studies.* 81(2): 608–650.

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

Boehmer, E., J. Musumeci, and A. B. Poulsen. 1991. "Event-study methodology under conditions of event-induced variance". *Journal of Financial Economics.* 30(2): 253–272.

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

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

Brown, S. J. and J. B. Warner. 1985. "Using Daily Stock Returns The Case of Event Studies". *Journal of Financial Economics*. 14: 3–31.

Carhart, M. M. 1997. "On Persistence in Mutual Fund Performance". *The Journal of Finance*. 52(1): 57–82.

Chandra, R., S. Moriarity, and G. Lee Willinger. 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(2): 398–408.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. "Double/debiased machine learning for treatment and structural parameters". *The Econometrics Journal*. 21(1): C1–C68.

Conley, T. G. and C. R. Taber. 2011. "Inference with "difference in differences" with a small number of policy changes". *Review of Economics and Statistics*. 93(1): 113–125.

Corrado, C. J. 1989. "A Nonparametric Test for Abnormal Security-Price Performance in Event Studies". *Journal of Financial Economics*. 23(2): 385–395.

Corrado, C. J. 2011. "Event studies: A methodology review". *Accounting and Finance*. 51(1): 207–234.

Dove, T., D. Heath, and J. B. Heaton. 2019. "Bias-Corrected Estimation of Price Impact in Securities Litigation". *American Law and Economics Review*. 21(1): 184–208.

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

Fama, E. F. and K. R. French. 1996. "The CAPM is Wanted, Dead or Alive". *The Journal of Finance*. LI(5).

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

Gelbach, J. B. and J. E. Fisch. 2021. "Power and Stasticial Significance in Securities Fraud Litigation". *Harvard Business Law Review*.

Gelbach, J. B. and J. R. Hawkins. 2020. "A Bayesian Approach to Event Studies for Securities Litigation". *Journal of Institutional and Theoretical Economics*. 176(1): 86–121.

Gelbach, J. B., E. Helland, and J. Klick. 2013. "Valid Inference in Single-Firm, Single-Event Studies". *Tech. rep.* No. 2. 495–541.

Haw, R. 2012. "Adversarial Economics in Antitrust Litigation: Losing Academic Consensus in the Battle of the Experts". *Vanderbilt Law Review*. 106(3).

Hein, S. E. and P. Westfall. 2004. "Improving Tests of Abnormal Returns by Bootstrapping the Multivariate Regression Model with Event Paraments". *Journal of Financial Econometrics*. 2(3): 451–471.

Imbens, G. W. and N. Doudchenko. 2016. "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis".

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer. 2015. "Prediction Policy Problems †". *American Economic Review: Papers & Proceedings*. 105(5): 491–495.

Kolari, J. W. and S. Pynnönen. 2010. "Event study testing with cross-sectional correlation of abnormal returns". *Review of Financial Studies*. 23(11): 3996–4025.

Kothari, S. P. and J. B. Warner. 2007. "Econometrics of Event Studies". In: *Handbook of Empirical Corporate Finance*. Vol. 1. 3–36.

Patell, J. 1976. "Corporate Forecasts of Earnings Per Share and Stock Price Behavior : Empirical Test". *Journal of Accounting Research*. 14(2): 246–276.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso". *Source: Journal of the Royal Statistical Society. Series B (Methodological)*. 58(1): 267–288.

Wooldridge, J. M. 2002. *Econometric analysis of cross section and panel data*. MIT Press.