

Valid Inference in Single-Firm, Single-Event Studies

Jonah B. Gelbach
Eller College of Management
University of Arizona

Eric Helland
Robert Day School
Claremont McKenna College

Jonathan Klick
University of Pennsylvania Law School

May 17, 2010

Abstract

We first discuss the role of single-firm event studies in law and finance scholarship and in securities litigation. We discuss the (previously known) invalidity of the standard, t -statistic-based approach to inference in single-firm event studies. We then use a broad cross-section of CRSP data for the 2000-2007 period to investigate the standard approach's performance using real-world data. Our results show that the standard approach is plagued by systematic, downward bias in asymptotic Type I error rates relative to desired significance levels. We then offer a very simple but statistically sound alternative, called the SQ test. We show analytically that the SQ test's asymptotic Type I error rate always equals the desired significance level. Using our CRSP data, we offer Monte Carlo evidence that in event studies with 100 pre-event observations, the SQ test performs very well at conventional significance levels. We then analyze the asymptotic power of the SQ test and the standard approach. The SQ test and the standard approach have the same size-corrected asymptotic power, which means that even when the standard approach is appropriate, there is no loss to using the SQ test. More relevant as an empirical matter, we show that the standard approach's downward bias in asymptotic Type I error rates brings severe power loss. Thus, previous studies of the impact of regulation, anti-trust rulings, corporate litigation and securities litigation on securities prices may suffer from systematic biases. By contrast, the SQ test has considerable asymptotic power, even against moderately sized fixed alternatives. We also show how to extend our methods to cases with multiple firms, and multiple events. Finally, we show that our SQ test is asymptotically equivalent to bootstrap procedures, including those evaluated in recent finance scholarship.

*Contact information: gelbach@email.arizona.edu, eric.helland@claremontmckenna.edu, and jklick@law.upenn.edu. We thank Ezra Friedman, Kei Hirano, Justin Wolfers, and participants at the 2007 Future of Securities Fraud Litigation conference at Claremont McKenna College, Northwestern University Law School, and the University of Pennsylvania Law School for helpful comments.

1 Introduction

Bhagat & Romano (2002, p. 141) open their article by writing that “Event studies are among the most successful uses of econometrics in policy analysis.” Fama (1991) attributes most of what we know empirically about corporate finance to event studies. Event studies have been used widely to examine proposed mergers, evaluate takeover policy, and assess the effects of a wide range of laws and regulations affecting corporations. Recent applications go beyond standard corporate finance topics. Two examples are Dellavigna & La Ferrara (Forthcoming), who use event studies as a forensic tool to assess whether firms violate United Nations weapons embargoes, and dellavigna:obama:efect:mar312010, who uses event studies to test for effects of Barack Obama’s electoral success on a variety of economic outcomes.

Single-firm event studies are less common in academic research than event studies that have multiple firms and/or dates. However, there are some important cases in the literature. One key example concerns the wealth effects of the Pennzoil-Texaco lawsuit. Cutler & Summers (1988) estimate that the Texaco case, which concerned tortious interference with a merger, destroyed more value than it created. Antitrust and litigation studies typically involve a small number of firms and a single event day.¹ Event study methodology is arguably even more important in the context of securities litigation. Since at least *Basic v. Levinson* (1988),² plaintiffs seeking to meet the reliance and materiality elements of a securities fraud case have been able to use event study evidence to establish that a security’s price movement was associated with a fraudulent statement made by a firm. Some courts have even effectively required the use of an event study for these purposes.³

Despite the wide use of event studies, the typical approach to event-study methodology turns out to be fundamentally flawed in important cases. The standard approach in testing for statistical significance of event effects involves comparing standard test statistics, like t -ratios, to critical values associated with the standard normal distribution. If firms’ abnormal returns come from a normal distribution, standard inference methodology is approximately justified. However, considerable evidence in the literature suggests that abnormal returns are in fact non-normal; we present extensive new evidence to this effect below.⁴ Alternatively, the standard approach is justified when there are many event dates, and for which the researcher is comfortable confining attention to average effects. In important applications, this is not true. For example, many, if not most, securities lawsuits involve only one or two event dates.

While this problem is most acute in the litigation setting, where analysis frequently centers on a single firm-event, it could affect inference in event studies more generally. For example, a single event might affect multiple firms, as considered in Greenstone, Oyer & Vissing-Jorgensen (2006) and Li, Pincus & Rego (2008). Alternatively, a series of relatively few events might affect the returns of relatively few firms. Bittlingmayer & Hazlett’s (2000) well-known study concerned the

¹Bittlingmayer & Hazlett’s (2000) study concerning antitrust enforcement and Microsoft is in a sense an exception that proves this rule. U.S. government enforcement actions against Microsoft occurred over many years. Microsoft’s key position in the software industry also raised the possibility that many firms’ values might be affected by such enforcement.

²485 U.S. 224.

³See, for example, *In re Oracle Securities Litigation*, 829 F. Supp. 1176, 1181 (N.D. Cal, 1993).

⁴For an early and classic reference on this point, see Brown & Warner (1985); Ford & Kline (2006) present a more recent discussion.

impact of a variety of antitrust enforcement events against Microsoft on relatively small portfolios of competitors. A case could certainly be made for allowing differential impacts of each event on each competitor separately.⁵

In section 2, we discuss the event study literature and the econometric literature related to the simple method for inference that we propose. In section 3, we briefly describe the data we use in this paper, which come from the Center for Research in Security Performance's (CRSP's) database for the years 2000 through 2007. In section 4, we introduce a variant of the typical statistical model used in event studies undertaken in securities lawsuits. We use analytical arguments to illustrate the importance of normality of the distribution of abnormal returns for achieving valid inference, even asymptotically.

In section 5, we present overwhelming evidence against the empirical plausibility of normality. We also present evidence of variation in key quantiles of securities' standardized abnormal returns distributions, which itself implies that the standard approach will yield Type I error rates different from the desired significance level. We then quantify the poor Type I error rate performance of the standard approach to testing for significant event-date effects. The standard approach's poor performance is systematically related to firms' abnormal returns distributions in a way that is both very easy to characterize and starkly demonstrated via graphical evidence.

In section 6, we discuss our alternative SQ test, which is based on sample quantiles of the empirical distribution of estimated abnormal returns. To motivate the test, consider the following important practical application of event studies. Suppose a firm discloses that its past quarterly earnings were substantially below the level claimed in an earlier earnings statement. A group of shareholders then files an action under SEC rule 10b-5, citing standard fraud-on-the-market arguments. Courts require a test of the null hypothesis that the corrective disclosure had no effect on a firm's stock price; the alternative hypothesis is that the event has reduced the value of the firm's stock.

To use our method, the witness would obtain data on the security's daily return and the market return for both the event date and a set of, say, $n = 100$ pre-event observations. She would then use OLS estimation to estimate the firm's beta and the coefficient on an event dummy, with the latter coefficient being the estimated event effect. All of these steps are taken in both the standard approach and in ours. To implement our test, the analyst would then calculate the fitted residuals from the estimated model, sort them, and find the 5th most negative value among the non-event dates. She would reject the null hypothesis if the coefficient on the event dummy were less than or equal to this value. Remarkably, the Type I error rate of this test converges to 0.05, regardless of the shape of the true distribution of abnormal returns.

The intuition for this result is simple. In a sample of 100 randomly drawn variables, the fifth-most negative element is the sample 0.05-quantile. It has long been known that sample quantiles are consistent estimators of population quantiles, so that the sample 0.05-quantile of a large collection of abnormal returns is an excellent estimate of the 0.05-quantile of the true underlying probability distribution for the abnormal returns. As we discuss below, this quantile is the key estimand in assessing whether a single event on a known date significantly reduced a firm's value. While the details below involve some technical points, this simple example illustrates

⁵For another example, Ford & Kline (2006).

how easy our sample quantile (SQ) test is to use in practice.⁶

In addition to describing the sample-quantile test in section 6, we also present empirical evidence from a Monte Carlo study of the SQ test's Type I error rate in samples with 100 pre-event dates. The results show that the SQ test performs very well with this sample size, which is not atypical of those used in event studies. We then turn to the issue of asymptotic power. We show analytically that on a size-corrected basis, the standard approach and our SQ test have the same asymptotic power. Given the standard approach's substantial asymptotic Type I error rate distortions, however, our estimates show that our SQ test has considerably greater power in the empirically relevant case when no size correction is made to the standard approach. Empirical results show that our test has considerable power even for moderate event-effect sizes. Interestingly, given the properties of observed data, the SQ test's power is greater, the more problematic is a firm's abnormal returns distributions for the standard approach.

In section 7, we discuss two potential extensions, involving multiple firms and multiple events. We discuss the relationship of our test to bootstrap procedures in the event-studies literature in section 8, and we conclude in section 9.

2 Relationship to Previous Literature

We begin this section by discussing the literature on event studies and corporate finance. We then discuss the relationship of the sample quantile test we propose to existing work in the statistical and econometric literature.

2.1 Event Studies

Event studies have been used in the academic literature to analyze many corporate finance issues, including the effects of earnings restatements and the adoption of various corporate governance mechanisms on firm value. They also play a prominent role in merger analysis and antitrust policy in both the academic and regulatory spheres. For example, law and economics scholars have used event studies to examine the effects of state-level legal changes, e.g., takeover statute enactments, as well as federal regulatory changes, e.g., Sarbanes-Oxley Act and the Private Securities Litigation Reform Act. Khotari & Warner (n.d.) offer an excellent recent review of the event-study literature, while Similarly, Campbell, Lo & MacKinlay (1997, p. 149) provide a very useful textbook discussion.

The popularity of event studies derives from their simple and elegant method of controlling for general market effects and other relevant covariates, thereby isolating causal effects of events

⁶One point worth noting, because it is sometimes misunderstood, is that sample quantiles do not “depend only on the observations no greater than them”. To be concrete, the sample .05-quantile in a size-100 sample does not depend only on five observations' values. It is true that the criterion for finding this sample quantile is to find the 5th smallest or most negative estimated abnormal return. However, there is no way to know which abnormal return is the fifth most negative without also knowing that 95 others are greater in value. In other words, sample quantiles generally depend on all n observations in a sample. This is a characteristic that sample quantiles share with the t -statistic used in the standard approach.

like a law’s passage, corporate governance adoption, and so on. Event-study methodology also provides a framework for determining whether estimated effects lie outside the range that could be expected due to ordinary random variation in stock returns, allowing researchers to determine whether the measured effect of an event is statistically significant.

While many practitioners have used the standard approach to inference in this context, several authors have recently noted and attempted to address the challenges of conducting inference with only a small number of events. Examples include Weinstein (2008), Klick & Sitkoff (2008), Hein & Westfall (2004) and Ford & Kline (2006). For example, in his study of American Express’s conversion to limited-liability status, Weinstein considers parametric alternatives to assuming normality of the abnormal-returns distribution. Klick and Sitkoff use a Monte Carlo, re-sampling approach in the spirit of permutation and bootstrap testing. Finally, Hein & Westfall (2004) and Ford & Kline (2006) use bootstrap re-sampling methods, which solves the standard approach’s asymptotic inference problems. We discuss the relationship between these studies and our own contributions in sections 2.2 and 8.

2.2 Related Statistical and Econometric Literature

In this section, we briefly discuss several strands of statistical and econometric research to which our SQ test is related: outlier detection, predictive tests of structural change, end-of-sample instability tests, permutation and randomization inference, and bootstrap-based inference.

As we discuss in section 4, when there is only one event, the estimated coefficient on the event-dummy in a market model equals the predicted residual from estimating the model without the event-date observation. The ratio of this coefficient estimate to the reported standard error is the usual t -statistic. In the literature on regression diagnostics, this ratio is known as the studentized residual and is often used as a measure of an observation’s influence, or leverage, in estimating slope coefficients. The purpose of evaluating leverage is typically to decide whether to omit an observation or otherwise address outlier influence for purposes of improving the performance of slope-coefficient estimation. In the litigation-event study context, this concern does not arise; the event date’s degree of leverage is interesting only for purposes of testing the size of the event effect.⁷

In the econometrics literature, the single-firm, single-event context can be thought of as a special case of a class of models considered by Taber & Conley (Forthcoming). Conley and Taber’s model involves a treatment group with a small number of cross-sectional observations and a large number of cross-sectional comparison-group observations. Our event date is analogous to their treatment group, which thus has exactly one observation, and our set of pre-event observations is analogous to their comparison group. Conley and Taber are primarily interested in estimating effects in different contexts from ours, like the effects of state-level policy reforms on labor market outcomes. However, their results can easily be shown to apply to our context, and a specialization of their Proposition 2 justifies our SQ test.⁸ We prove our results directly anyway, because their

⁷For a discussion of leverage and influence, see Belsley, Kuh & Welsch (2004).

⁸We became aware of this aspect of their work considerably after writing earlier drafts of this paper. An additional paper of which we became aware after writing earlier drafts of this paper is Simpson & Hosken (1998), who actually deploy the SQ test in an FTC working paper concerning one industry-specific empirical application, though without exploring its statistical properties. According to the FTC website, parts of this paper were

notation is substantially more developed than necessary for application to event studies like those we consider here.

Besides Taber & Conley (Forthcoming), the econometric literature most closely related to our test is the literature on structural change. An early and famous salvo in this literature is Chow (1960), who shows how to test the null hypothesis that the (linear) regression relationship between y and X for the next m_2 observations is the same as for the first m_1 observations, given the normality of regression residuals.⁹ Chow's focus was on testing whether the slope coefficients are the same in the two periods, however. He not only assumes normality of all residuals, he also assumes that the residual variances are the same in the two periods. In our context, this means that the Chow test imposes a zero event effect under both the null and alternative hypotheses. Exactly the opposite situation holds in the context that interests us: when conducting event studies, one typically maintains that the slope coefficients of the regression are the same and tests whether abnormal returns come from the same distribution before and after the event.

A variety of authors subsequently have explored the problem of testing for structural breaks when the break point is unknown; for example, see Andrews (1993). While this literature involves some similar statistical issues to our present context, the break point of interest is known in event studies. An early paper focusing on cases with known break points and allowing for non-normality is Dufour, Ghysels & Hall (1994). Dufour et al consider a more general econometric framework than ours, allowing for both multiple equations and nonlinearity, and lemma 3 below can be regarded as a special case of theirs. However, as Andrews (2003) has noted, their three approaches to estimating critical values all have disadvantages (namely, normality, asymptotic conservativeness, and the need to choose values of ancillary parameters).

The closest work to ours in the structural-change literature is Andrews (2003). Like ours, his basic test statistic involves predicting end-of-sample residuals. Also like ours, his critical values are estimated using the empirical distribution of predicted residuals from earlier in the sample. Translated into our terminology, Andrews's test allows an alternative hypothesis under which the event-date abnormal returns distribution differs from that for non-event dates. In fact, all of our results can be shown to follow by applying Andrews's proof argument to a suitable transformation of his test statistic. This transformation is necessary in the one-sided case because Andrews focuses on testing against two-sided alternative hypotheses. One-sided tests are of particular interest in important cases, like studies of litigation, antitrust or regulatory policy changes, since the relevant alternative hypothesis typically involves either a negative or positive event effect, but not both. For example, if a change in environmental regulation is associated with an increase in firm value, the researcher would not want to conclude that the new regulation increased compliance costs.

Another related literature involves permutation and randomization inference. Results in this literature rest on a key fact of which we make use below, that under the null hypothesis of no event effect, the event-date abnormal return comes from the same distribution as pre-event abnormal returns. For more on this literature, see Rosenbaum (2002).

A final related literature involves bootstrap-based inference. As we discuss below, our test can be shown to be a form of exact bootstrap test. Bootstrap test statistics and critical values are

subsequently published in Hosken & Simpson (2001) and Simpson (2001). Neither of those published papers appears to use the SQ test.

⁹Chow cites Mood (1950, pp. 304–05) as containing this result for the special case $m_2 = 1$.

computed by replacing an unknown distribution with an empirical distribution function (EDF) that consistently estimates the unknown distribution.¹⁰ In our context, we replace the event-date abnormal returns distribution with the EDF of predicted pre-event residuals, which is consistent under the null of no event effect.

The bootstrap paper most closely related to ours is Hein & Westfall (2004, HW), who evaluate bootstrap procedures proposed by Chou (2004), Hein, Westfall & Zhang (2001), and Kramer (2001). Like us, HW focus on inference in the single-event case, raising similar concerns to ours vis-a-vis the standard approach.¹¹ The tests HW evaluate yield are all based on Monte Carlo re-sampling algorithms. HW provide both solid heuristic arguments and Monte Carlo evidence that these procedures perform well in the single-firm case under both normal and some parametric non-normal data generating processes.

Relative to HW's excellent study, we make several contributions. First, while HW provide evidence on the S&P 500 and several insurance-based subindexes, we provide comprehensive empirical evidence for a wide array of individual securities. Second, while the bootstrap procedures that HW evaluate can be shown to yield valid inference as the number of re-samples grows, re-sampling methods can be cumbersome to explain and understand, especially for non-technical audiences. Third, we derive results that allow analytical comparisons, across the standard approach and our SQ test, of both estimated asymptotic Type I error rates and asymptotic power.¹² We do not mean to denigrate HW's excellent and revealing study; indeed, we believe it is a complement rather than a substitute to our own. However, for the reasons given above, we believe the results below represent a substantial contribution.

3 Data

We use data on securities returns from the widely used Center for Research in Security Performance (CRSP) database.¹³ In order to conduct the Monte Carlo study detailed below, we downloaded all daily observations from the CRSP database for the years 2000 through 2007. According to page 10 of the CRSP Data Description Guide, which is available at the Wharton Research Data Services website, the securities included in this sample include common stocks certificates, American Depositary Receipts (ADRs), shares of beneficial interest units (depository units, units of beneficial interest, units of limited partnership interest, depository receipts, etc.), closed-end

¹⁰The bootstrap idea was introduced by Efron (1979) as a generalization of jack-knife, or "leave-one-out" methods. Since then, an enormous literature has developed regarding a wide array of various testing procedures and many aspects of the bootstrap, including when it is known to succeed, when it is known to fail, and when it can improve on inference methods like the standard t -statistic approach considered above. Survey books include Davison & Hinkley (1997) and Efron & Tibshirani (1994); more technical but essential discussions for those interested in a deep understanding are available in Horowitz (2001) and Hall (1992).

¹¹Much of HW's interest lies in the single-event, multi-firm case. While our focus is primarily in the single-firm case, we do discuss the multi-firm case in section 7.1. Of course, HW's analytical results and arguments can be made comparable to our main discussion by setting the number of firms to one.

¹²HW offer analytical results on the asymptotic Type I error rate of the standard approach when the number of firms is large and all firms have the same abnormal returns distribution. Such results are interesting, but they offer no guidance in the single-firm case that interests us here.

¹³These data are available for academic use through the website of the University of Pennsylvania's Wharton Research Data Services.

mutual funds, foreigners on NYSE, AMEX, NASDAQ and NYSE Arca, Americus trust components (primes and scores), HOLDRs trusts, and REITs (real estate investment trusts).

Our initial data draw includes 14,587,459 daily observations. We kept only observations for which the daily returns (`ret`) and value-weighted returns including dividend (`vwretd`) variables were non-missing, eliminating 219,774 observations. We then drew, at random, five million observations.¹⁴ Of these observations, we kept only those associated with securities for which at least 500 daily observations remained; this criterion eliminated 303,746 observations. The resulting sample includes 4,696,254 daily returns observations on 3,050 securities, for an average of 1,540 observations per security.

We then calculated security-specific betas from a simple market model including a constant and the CRSP-provided value-weighted return, including dividends. The estimated beta for each firm is the OLS coefficient on the value-weighted return, with the firm’s daily return serving as the dependent variable. The daily fitted abnormal return is then the difference between the actual daily return and its predicted value based on the market model, as described in section 4 below.

The sample mean of fitted abnormal returns in our sample is 0 by construction. The sample standard deviation is 0.042, which means that shifting the distribution one standard deviation to the left entails a mean abnormal return of -4.2 percent of a security’s value. In much of our analysis, we standardize fitted abnormal returns by the standard deviation of firm-specific abnormal returns in our data. This standardization imposes mean-zero, standard deviation-one fitted abnormal returns at the firm level, facilitating comparisons both across firms and to normal distributions.

4 The Basic Framework With One Firm and One Event

We begin our discussion in section 4.1 by introducing the standard model for daily securities returns; throughout, we use the terms “firm”, “stock”, and “security” interchangeably. We focus in this section on the case in which there is a single firm and a single event (we discuss extensions to the multiple-firm and multiple-event cases in section 7). In section 4.2, we describe the standard approach to inference in event studies and detail the inference problem that plagues it in the one-firm, one-event case.

4.1 The Basic Framework

Daily event studies involve a security’s daily return, typically defined as

$$R_s \equiv \frac{P_s - P_{s-1}}{P_{s-1}}, \quad (1)$$

¹⁴We did all statistical work in version 10.1 of Stata using a 64-bit Dell machine running the x86.64 implementation of CentOS Linux. For the code we used to select data, we set the Stata seed to 99912, a value that was itself chosen at random using Stata’s `uniform()` random-number generator.

which is the proportionate change in the firm’s stock price, P , between days $s - 1$ and s .¹⁵ Event studies typically use the following model for firms’ daily returns:

$$R_s^j = X_s^j \beta^j + A_s^j, \quad (2)$$

where the superscript j indexes firms, the row-vector of data X_s^j includes 1 and a measure of the market return for day s and possibly other variables that might vary by firm; β^j is a vector of parameters that must be estimated; and A_s^j is firm j ’s day- s abnormal return, the component of the observed return that cannot be explained by X_s^j given the value of β^j . A common choice for the market return variable is the CRSP value-weighted return, which is what we use. Other variables sometimes included in so-called factor models are measures of firm size, the firm’s book-to-market equity, and momentum.¹⁶ For exposition, we focus on the simple market model here, so we include only the CRSP value-weighted portfolio as a non-constant regressor. For reference, Table 1 lists and defines the variables just described, as well as others introduced below. For the moment, we focus only on (2) as applied to a single firm, so we suppress the j superscript.

For concreteness, write the market return variable as R_s^m , so that $X_s = [1, R_s^m]$, and write the column vector of coefficients as $\beta = (\beta_0, \beta_m)'$. Thus, β_0 is the intercept, while β_m is the correlation of the firm’s return with the market return, commonly referred to as the firm’s beta. The key element in any event study has to do with the set of abnormal returns, $\{A_s\}$.

To account for the possibility of event effects, suppose that we have data on R_s and R_s^m for dates $s = 1, 2, \dots, n$. On date $e = n + 1$, an event occurs.¹⁷ Event examples of particular interest for litigation purposes in *Dura*’s wake are earnings re-statements and other corrective disclosures; merger announcements are another good example. To account for event effects, we re-write the day s abnormal return variable as $A_s = D_s \gamma + a_s$. Here, D_s is a dummy variable indicating whether day s is an event date: $D_s = 1(s = e)$, where the indicator function $1(\cdot)$ equals one when its argument is true and zero otherwise. The parameter γ is the true effect of an event on the level of the firm’s daily return. This effect could be either positive, negative, or zero. The a_s term represents the part of the abnormal return that is unrelated to the event; by definition, $A_s = a_s$ for all non-event dates. Thus, A_s can be viewed as a location shift of a_s , with the event effect γ being the shift parameter.

We follow common practice and assume that all abnormal returns for pre-event dates are *iid* conditional on the full set of regressors $\{X_s\}_{s=1}^{n+1}$ and come from the same distribution, which we name F_0 . Below we will allow these true abnormal returns distributions to vary across firms, in which case we call firm j ’s distribution F_0^j . While the *iid* assumption might seem strong, it is considerably weaker than the normality assumption required for the standard approach to deliver valid inference. In addition, results in Andrews (2003) and Taber & Conley (Forthcoming) show that our SQ test retains its good properties under a wide class of non-*iid* processes.¹⁸ Throughout, we will assume that abnormal returns are continuously distributed, so that F_0 is strictly increasing

¹⁵We have omitted notation concerning split factors and dividends from (1); the CRSP data we use do account for these factors.

¹⁶The three-factor model, which includes size and book-to-market variables in addition to the market return, was introduced by Fama & French (1992) and Fama & French (1993). Carhart (1997) added a momentum variable, resulting in the four-factor model.

¹⁷We could adapt this discussion to post-event data as well, but we stick with the pre-event case for clarity.

¹⁸Specifically, Andrews (2003) derives his results assuming only that $\{a_s\}$ is stationary and ergodic. This

on its entire support; again, this assumption is implied by but much weaker than normality. We refer to the standard deviation of F_0 as σ_a . Given the common assumption that events affect only the level of the daily return, it follows that a_e , the part of the event-date abnormal return that is unrelated to the event, also has distribution F_0 . In sum, for any y , we have¹⁹

$$R_s = X_s\beta + D_s\gamma + a_s, \quad Pr(a_s \leq y|\{X_s\}_{s=1}^{n+1}) = F_0(y), \quad s = 1, 2, \dots, n + 1. \quad (3)$$

Model (3) forms the basis for the so-called regression approach to estimating event effects. Using this approach, a researcher estimates β and γ jointly using ordinary least squares (OLS) estimation. She then evaluates the event's effect by testing the null hypothesis $H_0 : \gamma = 0$ against either a lower-tailed, two-sided, or upper-tailed alternative hypothesis

$$H_l : \gamma < 0 \qquad H_{two} : \gamma \neq 0 \qquad H_u : \gamma > 0. \quad (4)$$

In this paper, we focus on the lower-tailed case, in keeping with our example of testing whether a corrective disclosure reduces firm value. Analogous results for upper- and two-tailed alternatives follow as a matter of logic, so we will not treat them explicitly.

4.2 The Standard Approach to Inference

The standard approach to carrying out hypothesis tests involves estimating (3) by OLS and comparing the usual t -statistic for $\hat{\gamma}$ to critical values based on the standard normal distribution (or, occasionally, and more correctly, the Student's t distribution with the appropriate degrees of freedom). Let $\hat{\beta}$ and $\hat{\gamma}$ be the OLS estimates of β and γ . The t -statistic for testing the null hypothesis H_0 is

$$\hat{t} = \frac{\hat{\gamma}}{\widehat{\sigma}_\gamma},$$

assumption allows for some forms of both heteroskedasticity and dependence. Taber & Conley (Forthcoming) allow for dependence that satisfies strong mixing; given stationarity, strong mixing is a stronger assumption than ergodicity, though Conley and Taber's result follows generally under stationarity alone. An alternative approach would allow for the data generating process to evolve conditionally in a parametrically estimable way. For example, Weinstein (2008) allows for GARCH effects, together with a normality assumption on the white-noise part of the abnormal return. Given an algorithm for computing forecast variances, Weinstein's normality assumption could be further relaxed via quasi-maximum likelihood techniques; see Bollerslev & Wooldridge (1992) for more on QMLE as applied to GARCH models.

¹⁹For technical purposes, we will assume that $\{X_s, a_s\}_{s=1}^{n+1}$ is an iid sequence with $E[|X_s|]$, $E[X_s^2]$ and $E[|a_s|]$ all finite. This assumption ensures applicability of the weak law of large numbers to $n^{-1} \sum_{s=1}^n X'_s X_s$ and $n^{-1} \sum_{s=1}^n X'_s a_s$, which delivers consistency of $\hat{\beta}$ for β , below. The boundedness assumptions are reasonable given the nature of security returns; essentially they require only that the stock or market-portfolio price is bounded away from zero. The iid assumption is commonly made (implicitly) in event studies. It could be substantially weakened to allow for dependence; see White (2001, chapters 3-5) for details.

where $\hat{\sigma}_\gamma$ is the square-root of $\widehat{V}(\hat{\gamma})$, the estimated variance of $\hat{\gamma}$. The usual estimator of $\hat{\sigma}_\gamma$ is given by $\hat{\sigma}_\gamma^2 = \hat{\sigma}_a^2(D'M_xD)^{-1}$. The matrix M_x equals $I - X(X'X)^{-1}X'$, and the estimated standard error of the regression is $\hat{\sigma}_a^2 \equiv (n - 2)^{-1} \sum_{s=1}^{n+1} \hat{a}_s^2$, where $\hat{a}_s = R_s - X_s\hat{\beta} - D_s\hat{\gamma}$ is the fitted abnormal return for date s .²⁰ In conventional settings, a t -statistic like \hat{t} is well-behaved for one of two reasons:

- Case 1. If $\hat{\gamma}$ is exactly normally distributed, then under H_0 , \hat{t} has a Student's t distribution with degrees of freedom equal to $n - 2$. We use up one degree of freedom in estimating each of β_0 , β_m and γ . Since we have $n + 1$ observations, there are $n + 1 - 3 = n - 2$ degrees of freedom.
- Case 2. Suppose that $\hat{\gamma}$ is not exactly normal, but is instead only asymptotically root- n normal, so that the distribution of $\sqrt{n}(\hat{\gamma} - \gamma)$ converges in probability to $N(0, \sigma_\gamma^2)$ for some variance $\sigma_\gamma^2 > 0$. In this case, the finite-sample distribution of \hat{t} under H_0 is generally unknown. However, when n is large, the distribution of \hat{t} under H_0 is very well approximated by the standard normal distribution, given that $\hat{\sigma}_\gamma$ is a consistent estimate of σ_γ . As a consequence, it is common and generally appropriate to treat \hat{t} as if it were standard normal.

Under conventional conditions, then, a level- α test of H_0 would be based on comparing \hat{t} to the appropriate critical value based on quantiles of the standard normal distribution. Let Z be a standard normal random variable, and let z_α satisfy $Pr(Z \leq z_\alpha) = \alpha$, so that z_α is the α -quantile of the standard normal distribution. To test H_0 against the lower-tailed alternative H_l at level α , one would reject if and only if $\hat{t} \leq z_\alpha$. This is the standard approach. For reference, we state it as Procedure 1.

Procedure 1 (The Standard Approach to Inference).

1. Use OLS on the full sample of $n + 1$ observations to construct estimates $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\sigma}_\gamma$ of β , γ and σ_γ .
2. Construct $\hat{t} = \hat{\beta}/\hat{\sigma}_\gamma$.
3. Reject $H_0 : \gamma = 0$ against $H_l : \gamma < 0$ if and only if $\hat{t} \leq z_\alpha$, the α -quantile of the standard normal distribution (alternatively, use the α -quantile of the t_{n-2} distribution).

Unfortunately, a single-event study is not a conventional statistical context. The logic of the statistical argument for case 2 above rests on the applicability of a central limit theorem (CLT) to the behavior of the parameter $\hat{\gamma}$. CLT results typically hold in econometrics applications because (a) the statistic of interest, in this case $\hat{\gamma}$, can be written as a sample mean of a large number of observations, and (b) under extremely broad conditions, sample means are asymptotically normal. But when there is only one event date, $\hat{\gamma}$ cannot be written as a sample mean of many observations (the same result follows if separate effects are estimated for each of a set of multiple events). To sketch the explanation for this fact, the following lemma is helpful. We note that the lemma's first two parts are well known in the outlier-detection and predictive-test literatures²¹ and are also stated in Hein & Westfall (2004). The lemma's third part follows from the first two.

²⁰In a market model with $k > 3$ estimated parameters, one would replace $n - 2$ with $n + 1 - k$ in the formula above.

²¹For example, see Belsley et al. (2004) or Dufour (1980).

Lemma 1.

- (i) The event-date fitted abnormal return, $\hat{a}_e = R_e - X_e\hat{\beta} - \hat{\gamma}$, exactly equals 0, which implies $\hat{\gamma} = R_e - X_e\hat{\beta}$.
- (ii) The OLS estimate $\hat{\beta}$ from estimating (3) exactly equals the estimate that would be obtained by estimating (2) with the event date excluded.
- (iii) The estimated standard error of the regression, $\hat{\sigma}_a$, is the same regardless of whether we estimate equation (2) with the event date excluded or (3) with the event date included.

To show part (i), observe that the OLS estimation criterion is to choose g and b to minimize the sum of squared estimated residuals, $(R_e - X_e b - g)^2 + \sum_{s=1}^n (R_s - X_s b)^2$. Whatever value b takes, we can always make $\hat{a}_e^2 = 0$ by setting $\hat{\gamma}(b) = g = R_e - X_e b$. Thus $\hat{a}_e = 0$ is a necessary condition for minimization of the OLS criterion. Part (ii) then follows by observing that with \hat{a}_e always equal to 0, the OLS objective function for choosing b is the same when we estimate (3) with the event date included or estimate (2) with the event date excluded. To show part (iii), observe that because $\hat{\beta}$ is the same in each case, and the estimated event-date residual is zero, the sum of squared fitted residuals is identical in each case. Any difference would therefore have to come from the denominator used to estimate σ_a^2 . When we estimate (2) rather than (3), we add one observation but also one parameter, leaving the number of degrees of freedom unaffected at $n - 2$.

The next part of our argument concerns the distribution of the estimated event effect. With probability converging to one, this distribution becomes arbitrarily close to the distribution of γ plus the event-date abnormal return, a_e . Using (3) and part (i) of Lemma 1, we have $\hat{\gamma} = R_e - X_e\hat{\beta} = \gamma + a_e - X_e(\hat{\beta} - \beta)$. Since $\hat{\beta}$ is consistent for β , the probability limit of $\hat{\gamma} - \gamma - a_e$ equals 0. A basic result in statistics, called the asymptotic equivalence lemma by White (2001, Lemma 4.7, p. 67), holds that if $\text{plim}(\hat{\gamma} - \gamma) - a_e = 0$, then the asymptotic distribution of $(\hat{\gamma} - \gamma)$ and a_e must be the same. The distribution of a_e is obviously unaffected by the number of pre-event observations we choose to consider, so its asymptotic distribution is simply F_0 from (3). The asymptotic equivalence lemma thus tells us that $\lim_{n \rightarrow \infty} \text{Pr}(\hat{\gamma} - \gamma \leq y) = F_0(y)$. This result is fundamental: as long as we have a large number of non-event observations, the probability distribution of the estimated event effect under the null hypothesis of no event effect will converge to the true distribution of the event-date abnormal return. While this result can also be shown to hold as a special case of Corollary 3.1 of Dufour et al. (1994), we state it formally as a lemma for clarity.

Lemma 2.

$\text{Pr}(\hat{\gamma} - \gamma \leq y) \rightarrow F_0(y)$ as $n \rightarrow \infty$.

We next turn to the standard approach's asymptotic Type I error rate. It is possible to show that $\hat{\sigma}_\gamma$ converges in probability to σ_a , the standard deviation of the abnormal returns distribution F_0 .²² It then follows that $\hat{t} - \hat{\gamma}/\sigma_a$ converges to 0 in probability. Again using the

²²We have $\hat{\sigma}_\gamma^2 = \hat{\sigma}_a^2 (D' M_x D)^{-1}$. The term $D' M_x D = 1 - X_e (X' X)^{-1} X_e'$, whose second term equals $\text{trace}[X_e (X' X)^{-1} X_e'] = \text{trace}[(X' X/n)^{-1} X_e' X_e/n]$. By a law of large numbers, $(X' X/n)^{-1}$ converges in probability to its (finite) expectation, given moment conditions on X_s . Since X_e does not change with n , $X_e' X_e/n$ converges to zero. Therefore, the second term in $D' M_x D$ converges to 0 in probability, so the overall term converges to 1. Since $\hat{\sigma}_a^2$ is consistent for σ_a^2 , $\hat{\sigma}_\gamma^2 = \hat{\sigma}_a^2 (D' M_x D)^{-1}$ converges to σ_a^2 in probability, and thus $\hat{\sigma}_\gamma \xrightarrow{p} \sigma_a$.

asymptotic equivalence lemma, it follows that \hat{t} has the same asymptotic distribution as $\hat{\gamma}/\sigma_a$. We summarize results on the asymptotic distributions of $\hat{\gamma}$ and \hat{t} in the following lemma.

Lemma 3.

1. *The asymptotic distribution of the standard event effect estimate, $\hat{\gamma}$, is F_0 .*
2. *The asymptotic distribution of the t -ratio calculated using the standard approach satisfies $\lim_{n \rightarrow \infty} Pr(\hat{t} \leq y) = \lim_{n \rightarrow \infty} Pr(\hat{\gamma}/\sigma_a \leq y) = F_0(\sigma_a y)$.*

Lemma 3 is an important result linking the standard t -statistic, \hat{t} , to the distribution of the firm's abnormal return distribution under the null hypothesis of no event effect. When the true distribution of abnormal returns, F_0 , is a member of the normal family of probability distributions, $\hat{\gamma}$ is asymptotically normal, so $\hat{\gamma}/\sigma_a$ is asymptotically standard normal. Since $\hat{\sigma}_\gamma$ is consistent for σ_a , $\hat{t} = \hat{\gamma}/\hat{\sigma}_\gamma$ is also asymptotically standard normal when F_0 is a member of the normal family. Thus, $Pr(\hat{t} \leq z_\alpha)$ converges to α , so that the standard approach has asymptotically correct Type I error rate in this special case.

When F_0 is non-normal, though, the standard approach generally will lead to tests with asymptotic size errors. Worse still, the direction of the error is generally unknown without knowing the quantiles of the true distribution F_0 . As a consequence, the asymptotic error rate of the standard approach is unknown.

Using lemma 3, the most that can be said about the standard approach's Type I error rate in the general case of non-normal abnormal returns is that $Pr(\hat{t} \leq z_\alpha)$ converges to $F_0(\sigma_a z_\alpha)$. Define the α -quantile of the true abnormal returns distribution as y_α , so that $Pr(a_s \leq y_\alpha) = F_0(y_\alpha) = \alpha$ for $s \leq n$. When $y_\alpha < \sigma_a z_\alpha$, a lower-tailed test using the standard approach will reject more than $100 \times \alpha\%$ of the time, even in large samples. When the opposite holds, as we find below with many firms in our data, the standard approach will reject less than that percentage. Only when y_α happens to equal $\sigma_a z_\alpha$ will the standard approach reject at the desired Type I error rate α . We sum up these results with the following proposition concerning the asymptotic Type I error rate of the standard approach.

Proposition 1 (Type I Error Rates Using the Standard Approach).

The asymptotic Type I error rate of the standard approach is related to the desired level α as follows:

1. $F_0(\sigma_a z_\alpha) = \alpha$ when $y_\alpha = \sigma_a z_\alpha$, which is true if and only if F_0 is normal.
2. $F_0(\sigma_a z_\alpha) > \alpha$ when $y_\alpha < \sigma_a z_\alpha$.
3. $F_0(\sigma_a z_\alpha) < \alpha$ when $y_\alpha > \sigma_a z_\alpha$.

Figure 1 illustrates cases (b) and (c). In the top panel, we plot the density from a standard normal distribution together with the density from a Student's- t distribution with three degrees of freedom. The standard normal's 0.05-quantile is -1.645, while the $t(3)$ distribution's 0.05-quantile is -2.35. Thus, if abnormal returns follow a $t(3)$ distribution but one uses the -1.645 critical value for \hat{t} , one will reject the null hypothesis of no event effect more than five percent of the time.

In the figure’s bottom panel, we again plot the density from the standard normal distribution. The second density in this figure is estimated using standardized fitted abnormal returns from one of the firms in our sample.²³ To construct this density estimate, we first calculated \hat{a}_s for all dates s for which we have data for this firm. We then constructed standardized fitted abnormal returns as $\tilde{a}_s = \hat{a}_s / \hat{\sigma}_a$, where $\hat{\sigma}_a$ is the sample standard deviation of the 567 observations on \hat{a}_s that we have for this firm. If the firm’s true abnormal returns came from a normal distribution, then \tilde{a}_s would follow a Student’s- t distribution with degrees of freedom equal to the number of observations, minus 1. We have 567 observations on the firm in the picture, so the standard normal 0.05-quantile is a good approximation to the true 0.05-quantile of $\{\tilde{a}_s\}$ under normality of F_0 .²⁴ For the firm in the figure’s bottom panel, the sample-0.05 quantile was -0.999, considerably closer to the origin than the standard normal distribution value of -1.645. As the figure shows, using the standard normal critical value of -1.645 would yield a Type I error rate considerably below 0.05 for this firm.²⁵ Below, we provide evidence that this type of case is common for the significance levels we consider, 0.025, 0.05, and 0.10.

As a first pass at quantifying the importance of non-normality, we can estimate the asymptotic Type I error rate of the standard approach, under the assumption that all 3,050 securities’ returns come from the same (non-normal) distribution, i.e., $F_0^j = F_0$ for every $j = 1, 2, \dots, 3050$. Under this null hypothesis, the EDF of our pooled sample of abnormal returns, \hat{F}_{pooled} , is consistent for F_0 . As above, the standard approach’s true asymptotic Type I error rate for a level- α test against the lower-tailed alternative is $F_0(\sigma_a z_\alpha)$. We can estimate σ_a using the sample standard deviation of our abnormal returns sample, which is 0.042. Consider testing the null hypothesis of zero event effect against a lower-tailed alternative, at level $\alpha = 0.05$. Since $z_{.05} = -1.64$, and $0.042 \times -1.64 = -0.069$, the asymptotic Type 1 error rate for a level-0.05 test based on the standard approach is consistently estimated by $\hat{F}_{pooled}(-0.069)$.

Of the 4,696,254 abnormal returns in our pooled sample, fewer than 130,000 are less than or equal to -0.069. Our estimate of the Type I error rate based on a desired level-0.05 test using the standard approach on pooled data works out to roughly 0.027. In other words, given the assumption of a common abnormal returns distribution, the standard approach rejects only about half as often as the desired level of 0.05. Similar calculations show actual rejection rates of 0.019 and 0.044 for desired significance levels $\alpha = 0.025$ and $\alpha = 0.10$. These size distortions raise serious concerns for the standard approach. However, we arrived at them only after imposing the assumption that all 3,050 of the securities in our sample have the same underlying distribution of abnormal returns, i.e., $F_0^j = F_0$ for all $j = 1, 2, \dots, 3050$. In the next section, we test and reject this assumption. We then provide evidence that allows for firm-specific abnormal returns distributions.

²³We used Stata’s `kdensity` command to estimate this density function.

²⁴The 0.05-quantile of the $t(566)$ distribution is -1.6475502, compared to -1.6448536 for the standard normal distribution.

²⁵We chose the firm in this picture because its sample 0.05-quantile is especially close to the origin, allowing easy visualization of the source of under-rejection when using the standard approach. Among all firms in our sample: 99 percent of all firms in our sample had a sample-0.05 quantile of standardized fitted abnormal returns above this firm’s value of -0.999.

5 Non-Normality and Empirical Type I Error Rates Using the Standard Approach

In this section, we begin by testing three null hypotheses concerning the distribution of firms' abnormal returns. In section 5.1, we test whether all firms' abnormal returns come from the same normal distribution. We then test in section 5.2 whether they all come from the same non-normal distribution. In each case, we find overwhelming evidence against the null hypothesis. In section 5.3, we test whether each firm's abnormal returns come from a normal distribution, with the variances allowed to differ across firms. Again we reject easily. These results are important because they establish that Type I error rates for the standard approach must vary across firms. This fact implies that the standard approach will yield erroneous Type I error rates for at least some firms. We quantify the extent of these problems in section 5.4.

5.1 Is the Distribution of Pooled Abnormal Returns Normal?

We have stressed the critical reliance of the standard approach on normality of F_0 . We therefore begin our empirical discussion by considering whether the pooled sample of abnormal returns can plausibly be normal. A simple test of normality is the Jarque-Bera test based on two properties that hold for all normal distributions: they are symmetric, so that they have skewness equal to zero, and they have kurtosis equal to 3.²⁶ The pooled distribution of abnormal returns in our sample has sample skewness equal to 19.9 and sample kurtosis equal to 5,287, quite different from these values. The Jarque-Bera test statistic, JB ,²⁷ is based on the sample skewness and kurtosis values and is distributed χ^2 with two degrees of freedom under normality, with a level-.05 critical value of 5.99. The sample skewness and kurtosis values just listed yield a JB value in excess of five trillion. Clearly, then, the pooled distribution of abnormal returns is non-normal. We therefore reject the null hypothesis that all firms' abnormal returns come from the same normal distribution.

5.2 Do All Securities' Abnormal Returns Come from the Same Distribution?

Our next test concerns whether firms' abnormal returns come from the same (non-normal) distribution. Given the strength of the null hypothesis, that all 3,050 distributions are the same, one could devise a variety of tests. We focus on two tests that concern the behavior of the security-specific sample .05-quantiles, which we call $\hat{y}_{.05}^j$. We choose this statistic as the basis of our tests because variation in the true α -quantiles guarantees incorrect size for some securities, as discussed above.

It is of course possible that sample variation in key firm-specific sample α -quantiles is driven by random noise. To see how such a situation could occur, imagine that all securities' abnormal

²⁶A distribution's skewness is the ratio of its third central moment to the cube of its standard deviation. A distribution's kurtosis is the ratio of its fourth central moment to the fourth power of its standard deviation.

²⁷The JB statistic for distribution j equals $(n_j/6)(sk_j^2 + (1/4)(\kappa_j - 3)^2)$, where n_j is the sample size, sk_j is the sample skewness and κ_j is the sample kurtosis.

returns came from the same underlying distribution, F_0 . Given that we have data on 3,050 securities in our sample, we will wind up with 3,050 values of $\hat{y}_{.05}^j$ in any random sample drawn from this distribution. With so many draws, some of the security-specific sample .05-quantiles may appear to lie relatively far from the true population 0.05-quantile. Thus one needs a metric to evaluate whether the variation in sample .05-quantiles exceeds the variation to be expected under the null hypothesis of homogeneous abnormal returns distributions. We offer two approaches.

Our first approach is based on a permutation exercise. Under the null hypothesis that all firms' abnormal-returns distributions are the same, randomly re-ordering fitted abnormal returns across firm-day observations will not systematically affect the distribution of 3,050 firm-specific sample .05-quantiles, $\{\hat{y}_{.05}^j\}_{j=1}^{3050}$. Figure 2 plots two kernel density estimates of the cross-firm distribution of $\hat{y}_{.05}^j$. The first estimate, given by the solid line, is a density estimate for the sample 0.05-quantiles using the fitted abnormal returns based on the roughly five million actual returns we observe for our 3,050 firms. The second estimate, given by the dashed line, is a density estimate for the same statistics generated after randomly permuting all observed fitted abnormal returns across firm-day observations. If firms' abnormal returns came from the same underlying abnormal returns distribution, these two density estimates would be equal up to random variation induced by the permutation.

The figure suggests two important conclusions. First, much of the mass of the actual abnormal returns distribution of sample 0.05-quantiles lies considerably to the right of the permutation distribution. This means that firm-specific 0.05-quantiles are systematically closer to the origin than they would be if there were no heterogeneity in firms' abnormal returns distributions. As a consequence, the standard approach will reject less often at level 0.05 than it would in the absence of cross-firm distributional heterogeneity, even given the same pooled abnormal returns distribution. This result is consistent with the low rejection rate we found at the end of section 4.

Second, there is much more dispersion in the actual cross-firm distribution of sample 0.05-quantiles than in the permutation-based distribution. For example, the sample standard deviation of $\hat{y}_{.05}^j$ across j is 0.033 in the actual data, roughly an order of magnitude greater than the permutation distribution's 0.0036. This finding implies that different firms will have much more variation in rejection rates using the standard approach on any random sample than they would if firms had the same underlying abnormal returns distributions.

While the visual evidence in Figure 2 is overwhelming, it does not constitute a formal test. To carry one out, we use the fact that sample quantiles from continuous distributions are asymptotically normally distributed, regardless of the distribution generating them. Formally, if $F_0^j = F_0$ for all firms j , then $\sqrt{n_j}(\hat{y}_\alpha^j - y_\alpha^0) \xrightarrow{d} N(0, V_0)$, where y_α^0 is the true α -quantile of the common-across-firms abnormal returns distribution, $V_0 = \alpha(1 - \alpha)/[f_0(y_\alpha^0)]^2$, and f_0 is the density function associated with F_0 . Under the null, we can use a single estimate of the sample α -quantile of the pooled distribution, which is $\hat{y}_{.05}^0 = -0.050$. In other words, the sample 0.05-quantile of the pooled abnormal returns distribution is a reduction in firm value of 5 percent. We use Stata's `kdensity` command to estimate the pooled density at this value, with the result that $\hat{f}_0(\hat{y}_{.05}) = 1.72$.

Next, we define $\hat{Z}_{.05}^j = (n_j/\hat{V}_0)^{1/2}(\hat{y}_{.05}^j - \hat{y}_{.05}^0)$, with $\hat{V}_0 = 0.05 \times 0.95/1.72^2 = 0.016$. Under the null hypothesis of a common abnormal returns distribution, the sample $\{\hat{Z}_{.05}^j\}_{j=1}^{3050}$ must come from an approximately standard normal distribution. However, the actual sample mean of this distribution is 0.57, and the sample standard deviation is 8.93, strongly suggesting that the

underlying distribution does not have mean 0 and standard deviation 1. As with the permutation figure, these results show that the cross-firm distribution of sample 0.05-quantiles is shifted to the right and is much more dispersed than would be true if all abnormal returns came from the same distribution. In addition, the sample skewness of $\{\widehat{Z}^j\}$ is -1.07, and its sample kurtosis is 4.49, casting doubt on whether the distribution of sample 0.05-quantiles is normal at all. Indeed, the resulting JB statistic is 866, greatly exceeding the critical value of 5.99 for a level-.05 test. These results clearly reject the null hypothesis that all firms' abnormal returns come from a single distribution.

5.3 Are Security-Specific Abnormal Return Distributions Normal?

The preceding analysis shows that the pooled distribution of abnormal returns is not normally distributed, as well as that there is heterogeneity across firms in F_0^j , the underlying abnormal returns distributions. It remains possible that each security's abnormal returns distribution is normal, with variances differing across securities. This combination would cause both of the above results. It would also be unproblematic for the standard approach in event studies involving a single firm, since each abnormal return would come from some normal distribution.

To evaluate this possibility, we calculated the sample skewness and kurtosis values for abnormal returns within the size- n_j sample for each security j . All but two of the 3,050 securities in our sample have a value of $JB > 5.99$. Thus there is overwhelming evidence against firm-specific normality of abnormal returns. Combined with our theoretical results, the foregoing empirical findings suggest that the standard approach may involve substantial Type I distortions, when applied to firms one at a time.

5.4 How Much Does It Matter? Estimated Type I Error Rates Using the Standard Approach

Having established the non-normality of firm-specific abnormal returns distributions, we now assess its empirical implications for Type I error rate performance of the standard approach. In doing so, it will be convenient to standardize each \widehat{a}_s^j by $\widehat{\sigma}_a^j$, the sample standard deviation of abnormal returns for security j . The advantage of standardization is that it facilitates making direct comparisons between the standard approach and our results for our SQ test, below. Standardizing also entails no loss of generality, since it is an increasing transformation of each firm's set of abnormal returns, and probability distributions are invariant to increasing transformations. For notational simplicity, we will continue to use notation \widehat{a}_s^j when discussing standardized fitted abnormal returns.

Given that we work with standardized fitted abnormal returns, the standard approach has Type I error rate equal to $F_0^j(z_{.05})$. Thus the standard approach has the desired Type I error rate α for firm j only if the .05-quantile of standardized fitted abnormal returns $y_{.05}^j = z_{.05} = -1.64$. Except by accident, this equality will hold only when firm j 's abnormal returns distribution is normal, which we have seen is easy to reject for nearly all firms in our sample. Because we have seen compelling evidence that there is considerable variation in $y_{.05}^j$ across j , we expect the Type I performance of the standard approach to vary systematically with the true firm-specific values

of the sample 0.05-quantile $y_{0.05}^j$.

As above, let the firm- j sample of standardized fitted abnormal returns in our sample be $S_j \equiv \{\hat{a}_s^j\}_{s=1}^{n_j}$. Denote the EDF associated with each sample S_j as $\hat{F}_{n_j}^j(y)$. This function tells us the fraction of standardized fitted abnormal returns observations in the sample S_j whose value does not exceed the level y . Because an EDF is consistent for its underlying population counterpart, it follows that $\hat{F}_{n_j}^j(z_\alpha)$ is consistent for $F_0^j(z_\alpha)$, the true asymptotic Type I error rate of the standard approach for security j . To estimate the asymptotic size of the standard approach for each firm given $\alpha = 0.05$, we thus use $\hat{F}_{n_j}^j(-1.64)$, which is the EDF for firm j , evaluated at the standard normal distribution's 0.05-quantile. Our theoretical results imply that the estimated Type I error rate for the standard approach should be highest for values of $y_{0.05}^j$ that are far from the origin, i.e., very negative, lowest for values of $y_{0.05}^j$ close to the origin, and should fall as $y_{0.05}^j$ increases toward the origin.

In Figure 3, we plot the estimated asymptotic Type I error rates at three desired significance levels, $\alpha \in \{0.025, 0.05, 0.10\}$; these significance levels are indicated by horizontal lines. We indicate standard normal distribution quantiles corresponding to these levels using vertical lines; for example, the 0.025-quantile of the standard normal distribution is -1.96. The lighter, jagged lines plot the estimated asymptotic Type I error rate for the standard approach at each α . For firm j , our estimate of this error rate at level α is $\hat{F}_{n_j}^j(z_\alpha)$: this is the share of each firm's standardized fitted abnormal returns that fall below the standard normal distribution's α -quantile. The darker lines plot smoothed, nonparametric estimates of the average rejection rate at each value of the firm-specific sample α -quantile, which is given by the horizontal axis.²⁸

We note that for all three choices of α , we have included data on only those firms whose $\hat{y}_{0.05}^j$ lies in the middle 98 percent of the cross-firm distribution of \hat{y}_α^j ; this sample restriction avoids visual noise related to outliers. The left-most jagged series plots the estimated asymptotic Type I error rate for the standard approach at desired significance level $\alpha = 0.025$, with the middle and right-most series plotting the estimated asymptotic Type I error rates for $\alpha = 0.05$ and $\alpha = 0.10$.

The average value of $\hat{F}_{n_j}^j(z_{0.025})$ across our 3,050 firms is 0.023. This is close enough to the desired level of 0.025 that based on it alone, most researchers would conclude that the standard approach performs well for $\alpha = 0.025$. However, this conclusion would fail to account for cross-firm heterogeneity: correct asymptotic size for all firms would imply that the estimated asymptotic Type I error rates should form a horizontal line at the desired significance level α , up to sampling error. This expectation is treated rather roughly by Figure 3. Roughly a third of firms in the sample have $\hat{y}_{0.025}^j < -1.96$, so that their estimated asymptotic Type I error rate exceeds 0.025. The remaining two-thirds of securities have $\hat{y}_{0.025}^j > -1.96$, implying estimated asymptotic Type I error rates below $\alpha = 0.025$.

Turning to $\alpha = .05$, only three percent of firms have sample 0.05-quantile to the left of -1.64.

²⁸To compute these estimates, locally weighted scatterplot smoothing (lowess), available using Stata's `lowess` command. To estimate the conditional expectation of y given that $x = x_0$, lowess smoothing involves estimating a weighted least squares (WLS) regression of y_i on x_i , for all observations with x_i within a stated distance of x_0 . Thus, for each choice of x_0 , one estimates a separate WLS regression. For each such point, lowess requires the researcher to make three choices: the degree of the polynomial used in estimating each local regression, the weight function, and the share of the sample used. For the first two choices, we use a first-degree polynomial together with the tricube distance-weighting function, which are Stata's defaults. For the share of the sample included in each regression, Stata's default is 0.8; to avoid over-smoothing, we used 0.5, though we found results visually very similar to those computed using the default. For more details on lowess, see Cleveland (1979).

Thus, the standard approach would under-reject a true null of zero effect for the vast majority of firms. In fact, for a substantial share of firms the standard approach estimated asymptotic Type I error rate is below 0.025 when $\alpha = 0.05$. Indeed, the average value of $\widehat{F}_{n_j}^j(z_{.05})$ across our 3,050 firms is 0.036, which is a substantial average size distortion.

Next, consider $\alpha = 0.10$. Only one of our 3,050 firms has a sample 0.10-quantile below $z_{.10} = -1.28$, which explains why the entire upper series lies to the right of the vertical line at -1.28, since this part of the graph includes only the middle 98% of firms as measured by firm-specific sample 0.10-quantile. It is also remarkable that every firm in the graph has an estimated asymptotic Type I error rate below the desired level of 0.10. In addition, a substantial share of firms have estimated error rate below 0.05 when $\alpha = 0.10$, and the average value of $\widehat{F}_{n_j}^j(z_{.10})$ across our 3,050 firms is only 0.063. In sum, the evidence shows that the standard approach leads to substantial under-rejection for all three choices of α considered here.

An important feature of Figure 3 concerns the slopes of the smoothed estimated asymptotic Type I error rates. As we noted above, the graph of rejection rates against \widehat{y}_α^j for a procedure with correct size would be a horizontal line at α , up to sampling error. Not only do all three collections of error rates lines primarily lie below their desired α values, the error rates clearly fall as firms' sample α -quantiles rise toward zero. There is a simple reason why this happens. It is true by definition of sample quantiles that $F_0^j(y_\alpha^j) = \alpha$. The asymptotic error in using the standard approach is thus $\Delta^j(y_\alpha^j) = F_0^j(y_\alpha^j) - F_0^j(z_\alpha)$. Since distribution functions are non-decreasing, the asymptotic Type I error rate for the standard approach will equal, exceed, or be less than α whenever y_α is equal to, less than, or greater than z_α . Moreover, the derivative of Δ^j with respect to firm j 's α -quantile is $f_0^j(y_\alpha^j)$, which is strictly positive since it is a density function of a continuous random variable evaluated on the interior of its support. It follows that the error in using the standard approach is increasing in magnitude as y_α^j moves away from z_α .

5.5 Summary

In this section, we have presented overwhelming empirical evidence for the following conclusions:

1. Abnormal returns do not come from a single distribution, whether normal or of some other form.
2. There is systematic heterogeneity in firm-specific abnormal returns distributions.
3. Firm-specific abnormal returns distributions are non-normal for virtually all securities in our sample.
4. The standard approach to inference involves actual asymptotic size that is considerably different desired significance levels for three common choices of α : 0.025, 0.05, and 0.10. Moreover, these size distortions are systematically worse, the further is a firm's α -quantile from the standard-normal α -quantile. Empirically, our evidence shows that the vast majority of size distortions involve under-rejection of the null relative to the desired level α .

We now turn to our own approach.

6 The SQ Test

In this section, we explain and justify our alternative test for non-zero event effects. We then turn to Monte Carlo and other empirical evidence to assess the performance of this test.

6.1 Deriving and Characterizing the Test

In section 4 above, we explained why the estimated coefficient on the event dummy, $\hat{\gamma}$, converges in probability to $\gamma + a_e$, where we recall that γ is the true event effect on the firm's daily return, and a_e is the true event-date abnormal return (we drop the j superscript in this section for simplicity). As a consequence, $\hat{\gamma} - \gamma$ has the same asymptotic distribution, F_0 , as the event-date abnormal return, a_e . Lemma 3 thus shows that under $H_0 : \gamma = 0$, $\lim_{n \rightarrow \infty} Pr(\hat{\gamma} \leq y) = F_0(y) = Pr(a_e \leq y)$. Our test is based on this simple but powerful fact.

Suppose for a moment that we knew the value of the quantiles of F_0 . That is, for any α between 0 and 1, suppose we could determine y_α that satisfies

$$Pr(a_e \leq y_\alpha) = F_0(y_\alpha) = \alpha. \quad (5)$$

Since $\hat{\gamma}$'s asymptotic null distribution is F_0 , we could then use y_α as a critical value to test H_0 at level α : we would reject the null against the lower-tailed alternative whenever $\hat{\gamma} \leq y_\alpha$. The key challenge to asymptotically valid inference at level α is therefore to find a way to consistently estimate the α -quantile of F_0 . Under the assumptions commonly made in event studies, this is a surprisingly simple task, requiring only a trivial amount of additional work beyond that necessary for the standard approach. The following procedure characterizes our SQ test, which achieves asymptotic Type I error rate equal to α in testing H_0 against the lower-tailed alternative H_1 .

Procedure 2 (The SQ Test).

1. Estimate $\hat{\beta}$ and $\hat{\gamma}$ using OLS estimation of the market model (3).
2. For each non-event date $s \in \{1, 2, \dots, n\}$, calculate the fitted abnormal return $\hat{a}_s = R_s - X_s \hat{\beta}$.
3. Sort \hat{a}_s from least to greatest. Let the i^{th} order statistic be written $\hat{a}_{(i)}$, so that $\hat{a}_{(1)} \leq \hat{a}_{(2)} \leq \dots < \hat{a}_{(n)}$.
4. Next, define the greatest-integer operator $\lceil \cdot \rceil$ such that $\lceil x \rceil$ returns the integer c with the property that $x - 1 < c \leq x$. Define $c(\alpha, n) = \lceil \alpha \times n \rceil$, and find the $c(\alpha, n)$ order statistic of $\{\hat{a}_s\}$; call this value \hat{y}_α . In the case of $\alpha = 0.05$ and $n = 100$, we have $c(.05, 100) = 5$, so we find the 5th smallest (or most negative) value of \hat{a}_s . Call this value \hat{y}_α .
5. Reject H_0 against H_1 if and only if $\hat{\gamma} \leq \hat{y}_\alpha$.

It can be shown that as n grows, the probability that Procedure 1 rejects H_0 when it is true converges to α , which confirms that our test has asymptotically correct size. We now provide an intuitive explanation.

We have referred to the statistic \hat{y}_α as the $c(\alpha, n)$ order statistic of $\{\hat{a}_s\}$. This order statistic is also known as the sample α -quantile, a nomenclature we used repeatedly in our discussion of the standard approach above. To understand the logic of sample quantiles, it helps to define the EDF \hat{F} based on non-event date data:

$$\hat{F}(y) = \frac{1}{n} \sum_{s=1}^n 1(\hat{a}_s \leq y). \quad (6)$$

For arbitrary value y , the EDF \hat{F} tells us the share of all fitted abnormal returns that are no greater than y .²⁹ The sample α -quantile of $\{\hat{a}_s\}$ is defined as the most negative element of this set such that $\hat{F}(y) \geq \alpha$. Given the definition of $c(\alpha, n)$ above, \hat{y}_α is therefore both the sample α -quantile and the $c(\alpha, n)$ order statistic:

$$\hat{y}_\alpha \equiv \text{most negative element of } \{y : \hat{F}(y) \geq \alpha\} \quad (7)$$

$$= \hat{a}_{(c(\alpha, n))}. \quad (8)$$

Notice that \hat{F} involves sample, rather than “true”, information about abnormal returns in two ways. First, we define \hat{F} using n observations on fitted abnormal returns, rather than their true but unobserved counterparts: we use \hat{a}_s rather than a_s . Second, we have only a sample of n pre-event dates, rather than the entire population of abnormal returns values. One might think that the first issue is an important limitation.³⁰ However, it turns out that the estimated nature of \hat{a}_s is asymptotically irrelevant: $\hat{F}(y)$ is itself consistent for $F_0(y)$. To see why, define $F_n(y)$ as

$$F_n(y) = \frac{1}{n} \sum_{s=1}^n 1(a_s \leq y), \quad (9)$$

which is the EDF using true abnormal returns rather than fitted ones. It is not feasible to calculate $F_n(y)$, but it is still useful to imagine we could. While $F_n(y)$ is still only an estimate of the true population distribution F_0 , it is well known that F_n is uniformly consistent for F_0 (e.g., see van der Vaart (1998, Chapter 19)). What this means in our case is roughly that as n grows, $F_n(y) - F_0(y)$ converges to zero for all possible choices of y . Thus, if we could observe the true abnormal returns, we could estimate F_0 consistently using F_n .

The key theoretical fact justifying our feasible SQ test is that as n grows, the advantage in using F_n rather than \hat{F} vanishes. Because $\hat{\beta}$ is consistent for β , each \hat{a}_s is consistent for its true counterpart a_s , and this turns out to be enough to cause the difference between the feasible empirical distribution function $\hat{F}(y)$ and its population counterpart $F_0(y)$ to vanish. It is also true

²⁹Notice that \hat{F} is based on only the n non-event date observations on \hat{a}_s in a sample of size n that an analyst would have available for use in actual practice. This EDF should not be confused with the EDF $\hat{F}_{n_j}^j$ that we introduced above and work with below in estimating asymptotic Type I error rates; this latter object is the EDF for \hat{a}_s^j in our full CRSP sample.

³⁰For example, Hein & Westfall (2004, p. 465) observe that Monte Carlo approximation to the bootstrap distribution could be avoided if it were feasible to calculate F_n in (9).

that convergence of a cumulative distribution function and convergence of its associated quantiles are equivalent. Thus, since $\widehat{F}(y)$ converges to $F_0(y)$ under H_0 , the sample quantiles \widehat{y}_α must converge to the population quantiles y_α . It then follows that $\widehat{F}(\widehat{y}_\alpha) \xrightarrow{p} F_0(y_\alpha) = \alpha$, which we state formally a proposition for clarity.³¹

Proposition 2. *Under the assumptions above, as $n \rightarrow \infty$,*

1. $\lim_{n \rightarrow \infty} \widehat{F}(y) \rightarrow F_0(y)$ pointwise for all y in the support of F_0
2. $\lim_{n \rightarrow \infty} \widehat{y}_\alpha \rightarrow y_\alpha$
3. $\lim_{n \rightarrow \infty} Pr(\widehat{\gamma} \leq \widehat{y}_\alpha) \rightarrow \alpha$

Proposition 2 implies that the very simple Procedure 2 provides the basis for asymptotically valid inference, even though the procedure amounts to little more than sorting some fitted values. Unlike the standard approach of comparing t -statistics to critical values based on the standard normal distribution, the Type I error using Procedure 2 converges to the pre-specified level α as the number of non-event date observations grows. By contrast, the standard approach's actual Type I error rate generally differs from α in direction and magnitude that cannot be estimated without our or similar methods. While the Supreme Court's precedent in *Daubert* may make the standard approach inadmissible in court, our approach clearly satisfies *Daubert's* error-rate reliability criterion.

6.2 Monte Carlo Evidence on the Small-Sample Size of the SQ Test

We have already seen that the SQ test has asymptotically correct size, which is an improvement over the standard approach for firms with non-normal abnormal returns distributions. A natural next question concerns the small-sample behavior of the SQ test. It is always possible that an asymptotically justified method requires enormous sample sizes in practice. Event studies

³¹As noted above, the validity of Procedure 2 follows as a special case of Proposition 2 in Taber & Conley (Forthcoming). It could also be established by modifying details of the test statistic and critical values use in Andrews (2003). However, it is straightforward to prove the result directly, which we do in this footnote. Observe that $F_0(y) = Pr(a_s \leq y) = E[1(a_s \leq y)]$. Thus,

$$\widehat{F}(y) - F_0(y) = \frac{1}{n} \sum_{s=1}^n 1(\widehat{a}_s \leq y) - E[1(a \leq y)], \quad (\dagger)$$

where a is a generic draw from F_0 . Define the function $g(a_s, b) = 1(a_s - X_s(b - \beta) \leq y)$. Since g is an indicator function, it is completely bounded, and therefore it is bounded in expectation for any choice of b . Further, $g(a_s, b)$ is discontinuous only where $a_s = y + X_s(b - \beta)$. Since a_s has a continuous distribution by assumption, the probability that $g(a_s, b)$ is discontinuous at $b = \beta$ is zero. It follows that $g(a_s, b)$ is continuous at $b = \beta_0$ with probability one. Since a_s is iid by assumption and $\widehat{\beta} \xrightarrow{p} \beta$, Lemma 4.3 of Newey & McFadden (1994, p. 2156) implies that the right hand side of equation (\dagger) converges to zero in probability. This establishes convergence in probability of $\widehat{F}(y)$ to $F_0(y)$, pointwise for all y . Lemma 21.2 of van der Vaart (1998) asserts that such pointwise convergence is sufficient for the quantiles of \widehat{F} to converge to the quantiles of F_0 , which is sufficient for our procedure to have correct size. Non-iid data could be accommodated using the slightly more involved smoothing techniques Andrews (2003) uses to prove his Theorem 1, or by application of Proposition 2 of Taber & Conley (Forthcoming).

typically involve large but not huge samples. We thus present Monte Carlo evidence in this section for pre-event samples of size $n = 100$.

We investigate the SQ test’s small-sample performance with significance levels $\alpha = 0.025$, $\alpha = 0.05$, and $\alpha = 0.10$. For concreteness, we explain the procedure using $\alpha = 0.05$. Our Monte Carlo experiment consists of $m = 1, 2, \dots, 100$ repetitions of the following procedure:

1. For each firm j , we draw $n + 1 = 101$ values from the n_j observed values of (R_s^j, X_s) using random sampling with replacement. Observation s for firm j on Monte Carlo repetition m is $(R_{s,m}^j, X_{s,m})$, $s \in \{1, 2, \dots, n + 1\}$. For all firms, we set the “event” dummy $D_{s,m}$ equal to 1 for $s = n + 1$ and 0 for $1 < s \leq n$.
2. We estimate the model in (3). The fitted abnormal return for day $s \leq n$ in Monte Carlo iteration m is $\hat{a}_{s,m}^j = R_{s,m}^j - X_{s,m} \hat{\beta}^{j,m}$, where $\hat{\beta}^{j,m}$ is the OLS estimate of β . The estimated event effect is $\hat{\gamma}^{j,m} = R_{n+1,m}^j - X_{n+1,m} \hat{\beta}^{j,m}$, which also equald the coefficient on $D_{i,m}$, as stated in lemma 1.
3. We calculate the sample .05-quantile, $\hat{y}_{.05}^{j,m}$, of the $n = 100$ realizations of $\hat{a}_{s,m}^j$, i.e., the 5th order statistic of $\{\hat{a}_s^{j,m}\}_{s=1}^{100}$. We reject the null hypothesis against a lower-tailed alternative based on the SQ approach if and only if $\hat{\gamma}^{j,m} < \hat{y}_{.05}^{j,m}$. We define $r_{j,m} = 1$ if we reject for firm j on Monte Carlo iteration m , and $r_{j,m} = 0$ if we do not reject.

For each security j , we calculate the Monte Carlo rejection rate (MCRR) over the 100 repetitions of this experiment, i.e., $\bar{r}_j = 100^{-1} \sum_{m=1}^{100} r_{j,m}$. If a test has correct size for a given j , then it should exhibit a rejection rate of .05, up to Monte Carlo simulation error. While using only 100 Monte Carlo repetitions yields a very imprecise rejection rate for any given j , we can use two approaches to dealing with this imprecision.

First, we can investigate the grand mean Monte Carlo rejection rate when we pool over all 3,050 firms, $\bar{r} \equiv 3050^{-1} \sum_j \bar{r}_j$. For $\alpha \in \{.05, .10\}$, the Monte Carlo estimates of \bar{r} are 0.0503 and 0.1010, very close to the desired levels; we discuss the case of $\alpha = 0.025$ below. The nominal levels of 0.05 and 0.10 fall within 95% confidence intervals for these grand-mean rejection rates. Based on these pooled-across-firm results, 100 pre-event observations appears to be enough for the SQ test to deliver an actual Type I error rate that is indistinguishable from $\alpha \in \{.05, .10\}$ for practical purposes.

Our second approach is to again use local smoothing techniques to exploit the large number of securities in our sample, smoothing over the simulation error. This is a common approach to reducing the computational time needed to do reliable Monte Carlo studies with heterogeneity. As above, we use lowess smoothing.

For any choice of α , the asymptotic Type I error rate of our SQ test is always α .³² Thus we expect the graph of the MCRR to be horizontal at approximately level α , up to simulation error. Figure 4 plots lowess results for desired significance levels $\alpha \in \{0.025, 0.05, 0.10\}$. As with Figure 3 for the standard approach, we include horizontal lines at each desired significance level

³²As we discuss below, when $c(\alpha, n)$ is not an integer, we expect the SQ test’s size for given n to differ from its asymptotic size.

and vertical lines at the corresponding standard-normal quantiles. We do not plot the underlying, firm-specific Monte Carlo rejection rates, because as we noted above, they are individually very noisy with only 100 repetitions.

A first result displayed in Figure 4 is that the estimated Type I error rate in firm-specific samples of size-100 do not vary with firm-specific sample α -quantiles, except at the right tail of the graph for $\alpha = 0.05$ and the left tail of the graph for $\alpha = 0.10$. There are very few observations in these regions of the two graphs, e.g., roughly 5% of firms in our data have a 0.05-quantile to the right of -1.20, where the upward slope begins for the $\alpha = 0.05$ graph.³³ For the vast majority of firms, then, the SQ test’s rejection probabilities are independent of firm holds in typical sample sizes, as well as asymptotically. A second key result in Figure 4 is that the estimated Type I error rate is essentially indistinguishable from α for $\alpha \in \{0.05, 0.10\}$.

For $\alpha = 0.025$, Figure 4 suggests that the SQ test rejects slightly more frequently than the desired level. The raw mean rejection rate confirms this suggestion: the SQ test’s average Monte Carlo rejection rate across firms is 0.0303 at desired level 0.025, with asymptotic 95% confidence interval (0.0297, 0.0309). The fact that the test of desired level 0.025 actually rejects a true null with probability roughly 0.03 in a sample of size 100 has a simple explanation. In a sample of size 100, the relevant critical value is the $c(0.025, 100) = \lceil 2.5 \rceil$ order statistic. Since $\lceil 2.5 \rceil = 3$, we have $c(0.025, 100) = c(0.03, 100)$.

In other words, in any given sample of size 100, the SQ test uses the same critical value for a level-0.025 test as for a level-0.03 test. This property is not a defect of the test, but rather of the chosen sample size, 100. Given a desired significance level α , this example shows the importance of choosing n so that $c(\alpha, n)$ is an integer. When $c(\alpha, n)$ has a non-integer value, the SQ test’s small-sample performance will yield upward distortions in actual size by design.³⁴ To illustrate this point, we ran the same Monte Carlo procedure described above for $\alpha = 0.025$ and $n \in \{40, 80, 200\}$. We chose these pre-event sample sizes because each has the property that $c(0.025, n)$ is integer-valued: $c(0.025, 40) = 1$, $c(0.025, 80) = 2$, and $c(0.025, 200) = 5$. The average Monte Carlo rejection rate across firms was 0.0268 for $n = 40$, 0.0254 for $n = 80$, and 0.0252 for $n = 200$. The fact that the SQ test performs as well as it does with $n = 40$ is a testament to its robustness.

6.3 Asymptotic Power of the SQ Test

We now investigate the asymptotic power properties of our SQ test; we will also study asymptotic power for the standard approach. Many tests involving large samples are consistent, which means that they have asymptotic power equal to 1. For example, suppose we base a test on the null hypothesis that a firm’s beta is 0, $H_0 : \beta_m = 0$, when it is not actually 0. If we use a Wald test, then our test statistic is the square of the usual t -statistic, $n(\hat{\beta}_m)^2 / \hat{V}(\sqrt{n}\hat{\beta}_m)$. The probability of rejecting the null converges to 1 as the sample size grows, because $(\hat{\beta}_m)^2 \xrightarrow{p} (\beta_m)^2 > 0$, while the

³³Further, we believe the sloping tails in Figure 4 is likely an artifact of the way Stata’s lowest bandwidths are chosen near endpoints. This is a point we intend to investigate further.

³⁴An alternative approach in non-integer cases would be to interpolate between $c(\alpha, n)$ and $c(\alpha, n) - 1$ so as to smooth the rejection rate. This approach requires an interpolation algorithm, which could be difficult to choose in practice, given that abnormal returns exhibit non-normality of unknown form. We therefore do not pursue this point here.

denominator converges to a positive constant. Therefore, the Wald statistic diverges as $n \rightarrow \infty$, and the probability of rejecting goes to 1. Neither our SQ test nor the standard approach can have this property, even under normality of F_0 , since the number of event dates does not grow with n . Test consistency is simply not possible with a fixed number of events. Instead of consistency, then, we consider the more forgiving standard of test unbiasedness, which means that the test's power is greater than its size, i.e., Type I error rate.

6.4 Asymptotic Power of the SQ Test

Consider our SQ test's power against the lower-tailed alternative, $H_l : \gamma < 0$. This application arises when we are interested in testing whether a firm's corrective disclosure caused a reduction in its stock price, for example. Since γ is a scalar constant, $Pr(\hat{\gamma} \leq y) = Pr(\hat{\gamma} - \gamma \leq y - \gamma)$. Our SQ test's rejection probability given γ and level α is thus $Pr(\hat{\gamma} - \gamma \leq \hat{y}_\alpha - \gamma)$. Since \hat{y}_α is consistent for y_α , and since the asymptotic distribution of $\hat{\gamma} - \gamma$ is F_0 , it follows that $Pr(\hat{\gamma} - \gamma \leq \hat{y}_\alpha - \gamma)$ converges to $F_0(y_\alpha - \gamma)$. Therefore, our test's asymptotic rejection probability against a lower-tailed alternative is $F_0(y_\alpha - \gamma)$. Since $\gamma < 0$ when the alternative hypothesis is true, and since F_0 is strictly increasing given the assumption of continuously distributed abnormal returns, we have $F_0(y_\alpha - \gamma) > F_0(y_\alpha) = \alpha$. Therefore, our SQ test's asymptotic probability of rejecting the null against a true lower-tailed alternative is greater than its asymptotic size. This establishes unbiasedness.

6.5 Asymptotic Power of the Standard Approach

Unlike the SQ test, the standard approach need not be unbiased. Given that we work with standardized abnormal returns, so that $\sigma_a = 1$, the standard approach's asymptotic rejection rate given a true effect of γ is given by

$$Pr(\hat{\gamma} \leq z_\alpha) = Pr(\hat{\gamma} - \gamma \leq z_\alpha - \gamma),$$

which equals $F_0(z_\alpha - \gamma)$. This probability exceeds α only when $F_0(z_\alpha - \gamma) > F_0(y_\alpha)$. Since F_0 is strictly increasing, this inequality is violated when $y_\alpha > z_\alpha - \gamma$. The standard approach may yield either asymptotically biased or unbiased tests, depending on the magnitude of γ and the discrepancy between y_α and z_α . Biased tests will result for more firms when there are smaller true effects, i.e., γ closer to zero, and greater compression of the abnormal returns distribution relative to normality.

Since each test's rejection probability depends on F_0 , the relationship between the power of the standard approach and our SQ test will depend on the shape of the standardized abnormal returns distribution. Consider first the case when the standard approach has correct asymptotic size, so that $y_\alpha = z_\alpha$. In this case, the standard approach's asymptotic rejection rate for given γ is $F_0(y_\alpha - \gamma)$, exactly equal to the asymptotic power of our SQ approach. Thus when the standard approach provides correct asymptotic size, our SQ test entails no loss of asymptotic power relative

to the standard approach.³⁵

As we showed above, the standard approach does not generally have correct size as an empirical matter. Therefore, the fact that the two tests have equal size-corrected power is primarily of theoretical interest. As an empirical matter, what matters is whether the size distortions we documented above bring along especially low or high power. This point is especially relevant in the context of litigation and policy decisions. In each of these cases, the financial consequences of inference are potentially very large. Thus we will use our CRSP data to compare the estimated asymptotic rejection rate of the uncorrected standard approach to our SQ test’s estimated asymptotic rejection rate.

6.6 Empirical Estimates of Asymptotic Power

We use the security-specific values of \hat{y}_{α, n_j}^j to estimate our SQ test’s power for two event effect sizes, $\gamma = -1$ and $\gamma = -0.5$; the n_j subscript in \hat{y}_{α, n_j}^j indicates that this sample α -quantile is estimated using all n_j observations on firm j in our data. Since we continue to impose $\sigma_a = 1$ by working with standardized abnormal returns, these event effects correspond to one standard deviation and one-half standard deviation of each firm’s un-standardized abnormal returns distribution. The SQ test’s asymptotic power for security j is $F_0(y_\alpha^j + 1)$ with $\gamma = -1$, and with $\gamma = -0.5$, power is $F_0(y_\alpha^j + 0.5)$. Again letting $\hat{F}_{n_j}^j$ be the EDF of abnormal returns for security j based on all n_j available observations, we estimate these probabilities using $\hat{F}_{n_j}^j(\hat{y}_{\alpha, n_j}^j + 1)$ and $\hat{F}_{n_j}^j(\hat{y}_{\alpha, n_j}^j + 0.5)$. The results discussed above ensure that these probabilities are consistent for their population counterparts as n_j grows. Note that this evidence involves actual sample information only, rather than a Monte Carlo experiment.

In our sample, a one-half standard deviation drop in the standardized fitted abnormal returns corresponds to a 2.1 percent drop in firm value on the event date. This is not a particularly large effect, so it will not be surprising to find relatively low rejection rates. As a basis for comparison, suppose we knew that F_0 were normal, so that standardized abnormal returns had a standard normal distribution. In this case, we could compute the asymptotic power of the standard approach analytically at each choice of α , since then $Pr(\text{Reject}|\gamma, \alpha) = Pr(\hat{\gamma} \leq z_\alpha)$, which converges to $\Phi(z_\alpha - \gamma)$, where Φ is the standard normal cumulative distribution function. Notice that since the SQ test is asymptotically identical to the standard approach when F_0 is normal, the standard approach and the SQ test both have asymptotic power $\Phi(z_\alpha - \gamma)$ under normality of F_0 . We stress that since abnormal returns distributions are clearly non-normal, the

³⁵An additional implication of this result is that the two tests have the same size-corrected asymptotic power. When a test’s actual size differs from its desired level, power comparisons can be made to a test having correct size by correcting the size of the test of interest. This principle avoids privileging tests that reject too often, as can be seen by noting that one can achieve power equal to its theoretical maximum of one by always rejecting the null hypothesis. This “test” has actual size one, regardless of desired level α , a poor property to say the least. To correct for size distortions, one first replaces the critical value used in the test with a correct critical value, i.e. one that induces correct size. One then uses this correct critical value to calculate the test’s rejection rate against the specified alternative hypothesis (e.g., an effect size of one standard deviation). In the present context, the correct critical value for the standard test based on \hat{t} would be $y_\alpha/\hat{\sigma}_a$, i.e., this is the value that we would use in place of z_α . Since \hat{y}_α is consistent for y_α , a test based on comparing \hat{t} to $y_\alpha/\hat{\sigma}_a$ has the same asymptotic power as a test based on comparing $\hat{\gamma}$ to \hat{y}_α . Thus, our SQ approach and the size-corrected standard approach have identical asymptotic power.

only point of using the normality comparison is to fix a baseline level of power that could be considered reasonably attainable.

Table 2 reports asymptotic power for each choice of α and γ under normality of F_0 . It also reports estimated asymptotic power for the standard approach and for the SQ test. The standard approach's power given the empirical abnormal returns distributions is less than 70 percent of the power that would be achieved under normality in all but one case, when $\gamma = -1$ and $\alpha = .10$. By comparison, the SQ test's power exceeds that under normality in three of the six cases—greatly so in the case when $\gamma = -1$ and $\alpha = .10$. In the other three cases, the SQ test's power is generally close to its level under normality.

Leaving aside the issue of variation across firms' abnormal returns distributions, Table 2 shows that the SQ test has good asymptotic power even in highly non-normal data. By comparison, the standard approach consistently under-rejects the false null of no effect except with a relatively large event effect of one standard deviation and at the most forgiving significance level. Even then, it does much worse than the SQ test, with power roughly a third lower.

Figure 5 graphs estimated asymptotic power against firm-specific sample α -quantiles. The top graphs involve $\gamma = -0.5$, while the bottom ones involve $\gamma = -1$. The graphs on the left side are for the standard approach, while those on the right are for the SQ test. In each graph, we consider $\alpha \in \{0.025, 0.05, 0.10\}$. As in the size figures above, we plot the rejection rate on the vertical axis and the firm-specific sample α -quantile, $\hat{y}_{.05, n_j}^j$, on the horizontal axis. Horizontal lines represent the average rejection rate for each choice of α ; these averages appear in Table 2. We also continue to restrict attention to the middle 98 percent of firms as measured by sample α -quantile values. As above, the jagged, lighter series are firm-specific rejection rates, while the smoother series are lowest estimates of the rejection rate given the firm-specific sample α -quantile.

The results in Figure 5 illustrate two important facts in addition to those demonstrated in Table 2. First, holding constant the significance level α and the true effect size γ , the standard approach's power falls as the firm-specific α -quantile rises toward zero. Second, the opposite is true for the SQ test: power increases as the firm-specific α -quantile rises toward zero, holding constant α and γ .

There is a simple way to explain this combination of results. Recall that all abnormal returns have been standardized, so that the standard deviation is 1 for all j . Thus as the sample .05-quantile $\hat{y}_{.05, n_j}^j$ rises toward zero, the dispersion of the standardized abnormal returns distribution generally falls, i.e., the distribution becomes more compressed. As a general matter, then, a greater value of $\hat{y}_{.05, n_j}^j$ implies a greater mass between $\hat{F}_{n_j}^j(\hat{y}_{.05, n_j}^j)$ and $\hat{F}_{n_j}^j(\hat{y}_{.05, n_j}^j + \Delta)$ for positive Δ . This means that a fixed $\Delta = -\gamma$ will tend to move us further into the distribution when a security has a value of $\hat{y}_{.05, n_j}^j$ that is closer to zero. The result is a greater asymptotic rejection rate for the SQ test as we move to the right in the figures above.

Now consider the standard approach. Its critical value, z_α , does not increase as $\hat{y}_{.05, n_j}^j$ does: the standard approach uses the same critical value for all securities, regardless of their abnormal returns distributions. Therefore, its asymptotic rejection rate falls as we deal with firms with $\hat{y}_{.05, n_j}^j$ closer to zero. Thus the pattern shown in the graphs on the left side of Figure 5 is precisely what we should expect to see.

6.7 Summary of Asymptotic Power Results

Our power results are easy to summarize. First, the SQ test and standard approach have the same size-corrected asymptotic power. Second, the SQ test has substantial power, especially against an effect size as large as one standard deviation in magnitude. In several cases, the SQ test has greater power with actual data than either test would have under normality. Third, the SQ test's power increases with the degree of compression of a security's standardized abnormal returns distribution: the larger the departure from normality of the standardized abnormal returns distribution, the greater will be the SQ test's power. Third, the standard approach's asymptotic power is considerably lower than the SQ test's power for much of the range of the data. Fourth, the standard approach's asymptotic power decreases with the compression of a security's standardized abnormal returns distribution: the larger the departure from normality of the standardized abnormal returns distribution, the lower will be the power of the standard approach.

Substantively, our results for the standard approach using data from 2000-07 suggest the presence of a potentially severe bias against finding an event effect. Among other things, this suggests the potential for considerable anti-plaintiff bias in the context of recent securities litigation.

7 Extensions

We can extend the above results in numerous directions. The three most obvious involve multiple firms, multiple events, and the possibility of non-*iid* abnormal returns.

7.1 Multiple firms

One interesting extension involves the possibility of dealing with multiple firms experiencing an event on a single day. An interesting example is the effects on Microsoft's competitors of the June 7, 2000, order breaking up the company.³⁶ An additional, litigation-relevant example, could involve a firm that is sued by multiple competitors that all allege anti-competitive acts on a given day.

The SQ approach can be extended to the multiple-firm case by specifying (3) for each of $m > 1$ firms. As before, D_s^j is a dummy variable that equals 1 on the event date and 0 on all other dates. The parameter γ^j is firm j 's event effect. To implement the SQ approach, one estimates the m firm-specific equations, yielding $\hat{\gamma} = (\hat{\gamma}^1, \hat{\gamma}^2, \dots, \hat{\gamma}^m)$, the vector of the m firm-specific estimated event effects. Under the null hypothesis that all event effects are 0, the element of $\hat{\gamma}$ corresponding to firm j will have asymptotic marginal distribution F_0^j , the same as the firm's distribution of abnormal returns on non-event dates.

Working with $\hat{\gamma}$, or functions of it, requires that we derive its asymptotic distribution. Let $a_s = (a_s^1, a_s^2, \dots, a_s^m)$ be a random draw from the joint distribution of firms' abnormal returns. If firms' abnormal returns are mutually independent on any date s , then the asymptotic distribution of $\hat{\gamma}$ satisfies $F_{0a} = \times_{j=1}^m F_0^j$, i.e., the joint distribution is the product of marginals. When there is

³⁶See Bittlingmayer & Hazlett (2000) for more on Microsoft, antitrust enforcement, and event studies.

within-day, cross-firm dependence, this relationship does not hold, and we simply define the joint distribution of a_s as F_{0a} . In either case, it remains true that $\hat{\gamma} \xrightarrow{P} \gamma + a_e$. Therefore, $\hat{\gamma} - \gamma$ has asymptotic distribution equal to the asymptotic distribution of a_e , which is simply F_{0a} . Testing joint hypotheses involving multiple firms is thus a straightforward generalization of the single-firm case.

For example, let φ be some real-valued function of the vector of firm-specific abnormal returns that is bounded and continuous with probability one. Under the null hypothesis that the event has zero effect on all m firms, the asymptotic distribution of $\varphi(\hat{\gamma})$ is the distribution of $\varphi(a_s)$. This latter distribution can be estimated consistently using $\hat{F}_{0\varphi}(y) = (n+1)^{-1} \sum_{s=1}^n \varphi(\hat{a}_s)$, where \hat{a}_s is the vector of m fitted abnormal returns (standardized or not) for any non-event date. Our single-firm results on asymptotic size generalize easily to this case, though power properties must be established on a case-by-case basis.³⁷

One choice for φ would be $\varphi(a) = a' \hat{\Omega}^{-1} a$, where $\hat{\Omega}$ is a consistent estimate of Ω , the $m \times m$ variance matrix of a_s . This typical choice of φ would be appropriate only for two-sided alternatives, since it leads to tests that reject whenever $\hat{\gamma}^j$ is too far from zero in either direction, for some j . Moreover, this choice of φ for our SQ approach may not yield an unbiased test unless each a_s has a multivariate normal distribution. Under the null, this test has the same asymptotic power as the standard approach of using χ_m^2 critical values, which are unbiased under normality. However, we have not so far been able to establish that this choice of φ yields an unbiased test when normality does not hold.³⁸ Moreover, the resulting quadratic-form test is sensitive to all departures of $|\gamma^j|$ from 0, including those in the direction opposite to a one-tailed alternative of interest. This is an undesirable feature in testing against a one-sided alternative. For example, it could lead us to reject the null when a firm experiences a positive, rather than negative, event effect on the date of a corrective disclosure.

7.2 Multiple events

The case of multiple events is also easy to address. For simplicity we assume there is only one firm, since the previous section shows that it is easy to accommodate more than one firm, and

³⁷It is easy to establish that the generalized SQ test is unbiased whenever $\varphi(a) = G(c'a)$, where each element c_j of $c = (c_1, \dots, c_m)'$ is positive and G is continuous and strictly increasing. To see why, observe that the inter-day independence assumption implies $Pr(\varphi(a_e) \leq y) = Pr(\varphi(a_s) \leq y) = F_{0\varphi}(y)$ for all $s = 1, 2, \dots, n$. Since $\hat{\gamma} - \gamma \xrightarrow{P} a_e$, it follows that $Pr(\varphi(\hat{\gamma}) \leq y) = Pr(c'\hat{\gamma} \leq y) = Pr(c'(\hat{\gamma} - \gamma) \leq y - c'\gamma)$, which converges to $F_{0\varphi}(y - c'\gamma)$. Let y_α be the α -quantile of the distribution of $\varphi(a_s)$, so that $F_{0\varphi}(y_\alpha) = \alpha$. Since all elements of c are strictly positive, and since the lower-tailed alternative is $H_l : \gamma^j < 0$ for all j , under H_l we have $c'\gamma < 0$, and thus $y_\alpha - c'\gamma > y_\alpha$. It then follows from the fact that $F_{0\varphi}$ is strictly increasing that under H_l , $F_{0\varphi}(y_\alpha - c'\gamma) > F_{0\varphi}(y_\alpha) = \alpha$, which establishes that the test's power is greater than its significance level, establishing unbiasedness. An obvious choice for the j^{th} element of the column vector c is to use $c_j = \sigma_{a_j}^{-1}$, the inverse standard deviation of abnormal returns for the j^{th} security. This standard deviation is unknown, but it can be estimated from the pre-event abnormal returns. Using this choice of c has the effect of equalizing the scale of $c_j \hat{\gamma}^j$ across j , so that no subset of securities dominates the test statistic's distribution. Note that this test would perform less well for $H_l : \gamma^j < 0$ for some, rather than all, j , since firms with the negative effects could be masked by firms with positive effects, which is allowed under this choice of H_l .

³⁸The usual power results derived in standard textbooks lean heavily on the normality of $\hat{\gamma}$ under the alternative. Without normality, the quadratic form version of φ is not generally distributed χ^2 . Since $E[a_e] = 0$, it is easy to show that $E[\hat{\gamma}' \Omega^{-1} \hat{\gamma}]$ converges in probability to $a_e' \Omega^{-1} a_e + \gamma' \Omega^{-1} \gamma$, but this fact alone is insufficient to generate an analytical result on power without normality of a_e .

this way we can drop the firm super-script for clarity. With m event dates, the model is now $R_s = X_s\beta + \sum_{j=1}^m D_s^j\gamma^j + a_s$, where D_s^j is a date j -specific dummy variable and γ^j is the effect of the date- j event.

We focus here on the case in which $m = 2$, so that there are two event dates.³⁹ This case is of particular interest in the litigation context given some readings of Justice Breyer's opinion in *Dura*. That opinion states that plaintiffs must allege loss causation due to an initial fraudulent act (in the case of *Dura*, the alleged act was the company's claim that FDA approval was expected). This opinion suggests the possibility that both the date of the initial fraudulent act and the date of the corrective disclosure may require event analysis in future fraud-on-the-market cases.⁴⁰

Under the null hypothesis, $\gamma^1 = \gamma^2 = 0$. An alternative hypothesis relevant for a case like *Dura* would involve $\gamma^1 > 0 > \gamma^2$: on the date of the fraudulent act, the stock rises an unusually large amount, later to fall an unusually large amount on the date of the corrective disclosure. The arguments above establish that $(\hat{\gamma}^1 - \gamma^1) - a_{e_1} \xrightarrow{p} 0$ and $(\hat{\gamma}^2 - \gamma^2) - a_{e_2} \xrightarrow{p} 0$, where a_{e_j} is the abnormal return on the j^{th} event date. Under H_0 , $\gamma^j = 0$, so $\hat{\gamma}^j - a_{e_j} \xrightarrow{p} 0$. It then follows, as before, that the asymptotic null distribution of $\hat{\gamma}^j$ is $F_0(y) = Pr(a_s \leq y)$.

As noted above, it is common in event studies, especially those used in litigation, to assume conditional independence of returns across days. Under this assumption, the asymptotic joint distribution of $(\hat{\gamma}^1, \hat{\gamma}^2)$ is simply the product of marginals $F_{00}(y_1, y_2) = F_0(y_1)F_0(y_2)$. We can use this fact to construct a level- α test of the null $\gamma^1 = \gamma^2 = 0$ against the alternative that $\gamma^1 > 0 > \gamma^2$ as follows. Let y_{δ_1} and y_{δ_2} be the δ_1 - and δ_2 -quantiles that satisfy $F_0(y_{\delta_1}) = \delta_1$ and $F_0(y_{\delta_2}) = \delta_2$. Since $\hat{\gamma}^1$ and $\hat{\gamma}^2$ are asymptotically independent, it follows that

$$\lim_{n \rightarrow \infty} Pr(\hat{\gamma}^1 > y_{\delta_1} \text{ and } \hat{\gamma}^2 < y_{\delta_2}) = [1 - F_0(y_{\delta_1})]F_0(y_{\delta_2}) \quad (10)$$

$$= (1 - \delta_1)\delta_2. \quad (11)$$

For a level- α test, we choose δ_1 and δ_2 so that $\alpha = (1 - \delta_1)\delta_2$. A natural requirement is that in testing the joint null, we hold the two event effects to the same probabilistic standard, in which case $1 - \delta_1 = \delta_2 = \delta$. Thus, $\delta = \sqrt{\alpha}$. So, for a level- α test, define $\hat{y}_{\sqrt{\alpha}}$ and $\hat{y}_{1-\sqrt{\alpha}}$ as the sample $\sqrt{\alpha}$ - and $(1 - \sqrt{\alpha})$ -quantiles of the distribution of fitted abnormal returns. The rejection rule for our test is simple: reject the null that both event effects are 0 against the alternative hypothesis, of a positive date-1 and a negative date-2 effect, if and only if both $\hat{\gamma}^1 > \hat{y}_{1-\sqrt{\alpha}}$ and $\hat{\gamma}^2 < \hat{y}_{\sqrt{\alpha}}$. This makes intuitive sense: we reject the joint null whenever the two event dates' estimated effects are simultaneously far from the origin in the directions that the alternative hypothesis specifies.

It is interesting to note that this result is very different from the naive approach of conducting two separate level- α tests, one for each of $\hat{\gamma}^1$ and $\hat{\gamma}^2$, a point we have made elsewhere.⁴¹ To illustrate, suppose that F_0 were actually standard normal. In that case, the critical values for a level- α test would be 0.76 for $\hat{\gamma}^1$ and -0.76 for $\hat{\gamma}^2$: any time $\hat{\gamma}^1 > 0.76$ and $\hat{\gamma}^2 < -0.76$ both

³⁹It is straightforward to generalize to more dates. We consider only the two-date case both for brevity and because of the obvious application to securities litigation explained here.

⁴⁰For example, this is the reading favored by Dunbar & Mayer (2006), whose analysis would generally require an event study assessing the effects of both the initial alleged mis-statement and the subsequent corrective disclosure.

⁴¹See Gelbach, Helland & Klick (2009) for a discussion of this issue.

occur, we would reject the null at level 0.05. By contrast, the naive approach would reject only if $\hat{\gamma}^1 > 1.64$ and $\hat{\gamma}^2 < -1.64$. Given that F_0 is standard normal, the probability of this joint event is only $0.05^2 = .0025$. Thus the naive approach would radically under-reject the null. Requiring such a test is tantamount to changing the rules of the litigation game against plaintiffs. Justice Breyer’s opinion in *Dura* says nothing of changing the standard, i.e., the significance level, needed for proof. Rather, it concerns the facts that must be established for any given standard of proof. We believe it would be mistaken to implicitly read a new, much more stringent, standard of proof into a decision that is entirely silent on the point.

Finally, we note that exactly the opposite issue arises if the alternative hypothesis is that only one of the two event-date abnormal returns is non-zero. In this case, we must choose critical values that account for the fact that we get two draws from F_0 under the null. The typical approach in the statistics literature is to use the Bonferonni correction for multiple draws, which involves using test-specific significance levels $\delta = \alpha/2$, or 0.025 in the level-.05 case.⁴²

8 Relationship to the Bootstrap

In this section, we briefly explain why the method proposed above can be thought of as a bootstrap method of inference. The discussion in this section is somewhat technical. However, this discussion is not necessary to understand the logic of our test. Our purpose in this section is to explain why bootstrap methods like those evaluated in Hein & Westfall (2004) are asymptotically equivalent to ours, despite the relative simplicity of the SQ test.⁴³

Consider some test statistic T_n computed from the observed data. For a lower-tailed hypothesis, testing requires determining a critical value $t_{crit} = T_{crit}(\alpha)$ such that one will reject at level α if and only if $T_n \leq t_{crit}$. In our context, two obvious choices of T_n are the estimated coefficient on the event dummy, $\hat{\gamma}$, and its standardization, $\hat{t} = \hat{\gamma}/\hat{\sigma}_a$. Consider $\hat{\gamma}$ first. To make clear the dependence of $\hat{\gamma}$ on the observed sample, write

$$\hat{\gamma} = \hat{\gamma}_n(\{R_s, X_s\}_{s=1}^{n+1}) \tag{12}$$

$$= R_{n+1} - X_{n+1}\hat{\beta}_n(\{R_s, X_s\}_{s=1}^n), \tag{13}$$

where $\hat{\beta}_n = (\sum_{s=1}^n X'_s X_s)^{-1}(\sum_{s=1}^n X'_s R_s)$ is the OLS estimate from the size- n pre-event sample, as above.

From this discussion, we see that $\hat{\gamma}$ depends on the entire sample $\mathbb{S}_{n+1} = \{R_s, X_s\}_{s=1}^{n+1}$. It follows that the probability distribution of $\hat{\gamma}$ depends on the population joint distribution generating \mathbb{S}_{n+1} .

⁴²Because they ignore the possibility that both date-specific test statistics might cause the test to reject, Bonferonni corrections are slightly conservative. The exact value for δ is 0.0253 in the case $m = 2$. In general this will make little difference asymptotically, though with relatively small n , it could matter slightly in practice. For instance, if $n = 120$, then $.025n = 2.975$, so we would use the third order statistic of \hat{F} as our estimated critical value, while $.0253n = 3.0107$, so we would use the fourth order statistic.

⁴³Much of the notation in this section follows Horowitz (2001), which provides an extensive discussion of bootstrap theory.

Under our iid assumptions, this joint distribution is just the $(n+1)$ -fold product of the population joint distribution of $\{R_s, X_s\}$ for any s . Since daily firm returns and firm abnormal returns are deterministically related given X_s , we can instead work with the population joint distribution of $\{X_s, a_s\}$, which we call J_0 . Similarly, since $\hat{\sigma}_a$ depends on the same sample that $\hat{\gamma}$ does, the distribution of \hat{t} is also completely determined by J_0 , given the pre-event sample size n .

For given choice of $T_n \in \{\hat{\gamma}, \hat{t}\}$, we can thus write the test statistic's probability distribution from a sample with n pre-event observations as $Pr(T_n \leq y) = G_n(y|J_0)$. The n subscript in this notation indicates the sample size on which the probability distribution for $\hat{\gamma}$ is based. The conditioned distribution J_0 indicates the probability law that generates our data (we leave implicit that the size- n law is $\times_{s=1}^{n+1} J_0$). Letting J_{0x} be the marginal distribution of X_s , we thus have⁴⁴ $Pr(X_s \leq x, a_s \leq a) = J_0(x, a) = J_{0x}(x)F_0(a)$. We can now be more concrete about the distribution of the test statistic T_n : $Pr(T_n \leq y) = G_n(y|J_{0x}, F_0)$.

If we could calculate the quantiles of the distribution $G_n(\cdot|J_{0x}, F_0)$, we would be able to determine its α -quantile, $y_{\alpha, n}$. This would then serve as our critical value for a level- α test, since by definition of $y_{\alpha, n}$, $Pr(T_n \leq y_{\alpha, n}) = G_n(y_{\alpha, n}|J_{0x}, F_0) = \alpha$. The n subscript on $y_{\alpha, n}$ shows that the α -quantile of G_n varies with n , which is a key part of the challenge of valid inference. For practical purposes, knowing $y_{\alpha, n}$ for all α and knowing G_n are equivalent, so we can focus on whichever of these objects is more convenient in our discussion.

In some cases, G_n is known. For example, suppose that $T_n = \hat{t}$, and F_0 is normal. We have already seen that \hat{t} has a Student's- t distribution with $n-2$ degrees of freedom in this case. Thus when F_0 is normal, $G_n(\cdot|J_{0x}, F_0) = t_{n-2}$. The key fact here is that t_{n-2} depends on neither J_{0x} nor F_0 ; we say \hat{t} is pivotal, in such a case. Provided that F_0 is a member of the normal family, it doesn't matter which member it is, and it doesn't matter what form J_{0x} takes: knowing that F_0 is normal is sufficient for exact knowledge of the distribution $G_n(\cdot|J_{0x}, F_0)$. To be concrete, let $N(0, \sigma_a^2)$ be the normal distribution with mean zero and variance σ_a^2 . When $T_n = \hat{t}$, the standard approach can be viewed as replacing $G_n(\cdot|J_{0x}, F_0)$ with $G_n(\cdot|J_{0x}, N(0, \sigma_a^2)) = t_{n-2}$. We know that this is exactly correct when F_0 is indeed normal, and not otherwise, because

$$\lim_{n \rightarrow \infty} t_{n-2} = \Phi, \quad \text{while} \quad \lim_{n \rightarrow \infty} G_n(y|J_{0x}, F_0) = F_0(y/\sigma_a). \quad (14)$$

This discussion provides another way of understanding the problem with the standard approach under non-normality: $Pr(\hat{t} \leq y/\sigma_a)$ converges to a distribution that has variance one, but isn't generally normal. Using standard normal critical values will thus entail true asymptotic Type I error rates that generally differ from the desired level.

In the general case when test statistics are not pivotal, the distribution G_n will be unknown. Finding useable critical values thus requires that the quantiles of G_n must be estimated. Stepping outside the particular context of event studies, the most common approach to estimating a test statistic's distribution is to use large-sample theory to show that

⁴⁴When X_s is vector-valued, we use the convention that $X_s \leq x$ means that each element of X_s is no greater than the corresponding element of x .

$$\lim_{n \rightarrow \infty} G_n(y|J_0) = G_\infty(y). \quad (15)$$

Typically, the limiting distribution G_∞ is standard normal. For example, this would be the case if we used a t -test to test the null hypothesis that a firm's beta, β_m , equaled zero. As the sample size used to compute $\hat{\beta}_m$ grows, the null distribution of the statistic $\hat{t}_\beta = \hat{\beta}_m / \hat{\sigma}_\beta$ converges to the standard normal distribution, regardless of the form of the distribution J_0 (provided it satisfies assumptions sufficient to apply relevant laws of large numbers and central limit theorems, of course). In this case, then, $\lim_{n \rightarrow \infty} G_n(y|J_0) = \Phi(y)$, where Φ is the standard normal cdf. The key aspect of a result like (15), though, is not standard normality per se, but rather that the distribution G_∞ does not depend on J_0 . A test statistic T_n whose limiting distribution has this property, of being entirely known, is called asymptotically pivotal.

The asymptotic-theory approach to inference involves replacing $G_n(\cdot|J_0)$ with $G_\infty(\cdot)$. One then uses as critical values quantiles of G_∞ , rather than quantiles of G_n , i.e., $y_{\alpha,\infty}$ rather than $y_{\alpha,n}$. Provided that n is large, asymptotic theory will yield very similar inference to what one would get if the quantiles of G_n were known exactly. This fact is the foundation on which most modern econometric inference is based. However, in the present case, we have already seen that G_∞ involves the distribution for a single day's abnormal return, F_0 , which is unknown except by assumption.

An alternative approach to asymptotic theory is to use the bootstrap. Bootstrap-based inference on a test statistic T_n proceeds by replacing J_0 with some alternative distribution, \hat{J} . The resulting distribution can be written $G_n(\cdot|\hat{J})$, so that $Pr(T_n \leq y|\hat{J}) = G_n(\cdot|\hat{J})$. To implement the bootstrap, then, one needs two things: a good choice of \hat{J} , and a method to calculate $G_n(y|\hat{J})$ for each y . "Good", in this case, primarily means that \hat{J} is consistent for J_0 , so that as n grows, the distribution \hat{J} becomes arbitrarily close to the true population distribution J_0 .

Given a choice of \hat{J} , it is typically not possible to calculate $G_n(y|\hat{J})$ analytically. This is where the idea of Monte Carlo re-sampling comes in. Because \hat{J} is known, we can sample independently from it $n + 1$ times, with replacement, and use the resulting sample to calculate an analogue of the test statistic T_n . Calling the b^{th} instance of this analogue T_{nb} , we repeat this sampling B times, which yields a sample of B observations, $\{T_{nb}\}_{b=1}^B$, on the test statistic when J_0 is replaced by \hat{J} . As $B \rightarrow \infty$, the EDF of this sample can be shown to converge to the distribution $G_n(\cdot|\hat{J})$.

Usual choices of \hat{J} involve empirical distribution functions. One approach, known as the non-parametric or pairs bootstrap, is to put probability $1/(n + 1)$ on each $\{R_s, X_s\}$ pair in the observed sample $\{R_s, X_s\}_{s=1}^{n+1}$. Another approach, that taken by Hein & Westfall (2004), is to use the observed value of X_s for each s but approximate the distribution G_n under sampling from the EDF of fitted abnormal returns. Such approaches are known as residual bootstraps. The following procedure is an example of a typical residual bootstrap procedure that could be used to test the null hypothesis of $\gamma = 0$ against a lower-tailed alternative at level α :

1. Use OLS to estimate $\hat{\beta}$, $\hat{\gamma}$, $\hat{\sigma}_\gamma$ and $\hat{t} = \hat{\gamma} / \hat{\sigma}_\gamma$ using the observed sample of $n + 1$ observations.
2. For each $s = \{1, 2, \dots, n + 1\}$, calculate the fitted abnormal return $\hat{a}_s = R_s - X_s \hat{\beta}$. Note that the EDF of these values is \hat{F} from above.

3. Repeat the following B times, with iteration indexed by b :

- (a) Draw $n + 1$ times from the EDF of fitted abnormal returns, \widehat{F} , independently and with replacement. Call the s^{th} draw a_{sb}^* .
- (b) For each $s = \{1, 2, \dots, n + 1\}$, define $R_{sb}^* = X_s \widehat{\beta} + a_{sb}^*$, where $\widehat{\beta}$ is the OLS estimate from step 1 above.
- (c) Re-estimate the market model from step 1 using the re-sample $\{R_{sb}^*, X_s\}_{s=1}^{n+1}$. Call the resulting estimates of the event dummy's coefficient and its standard error, γ_b^* and $\sigma_{\gamma,b}^*$, and define $t_b^* = \gamma_b^* / \sigma_{\gamma,b}^*$.

4. Find the sample α -quantile, $y_{\alpha,B}^*$, of the re-sampled t -statistics, $\{t_b^*\}_{b=1}^B$. Reject H_0 if and only if \widehat{t} from step 1 is less than $y_{\alpha,B}^*$.

This procedure can be shown to yield asymptotically valid inference, even though there is just one event. First, observe that consistency of sample quantiles for population quantiles implies that $y_{\alpha,B}^*$ in step 4 is consistent for the α -quantile of the distribution of t_b^* as $B \rightarrow \infty$. Thus we need only show that t_b^* and \widehat{t} have the same asymptotic distribution as $n \rightarrow \infty$. From step 3b, we have $R_{sb}^* = X_s \widehat{\beta} + a_{sb}^*$. By lemma 1, $\gamma_b^* = R_{n+1,b}^* - X_{n+1} \beta_b^*$, where β_b^* is the estimated coefficient on X_s from step 3c. Arguments above imply that $\gamma_b^* = a_{n+1,b}^* - X_{n+1}(\beta_b^* - \widehat{\beta})$. As n grows, $\beta_b^* - \widehat{\beta} \xrightarrow{p} 0$ for each bootstrap iteration b . Therefore, $\text{plim}_{n \rightarrow \infty} \gamma_b^* - a_{n+1,b}^* = 0$, so that γ_b^* and $a_{n+1,b}^*$ have the same asymptotic distribution as n grows. From step 3a, we know that $a_{n+1,b}^*$ was drawn from the EDF of fitted abnormal returns. Thus, $\lim_{n \rightarrow \infty} \text{Pr}(a_{n+1,b}^* \leq y) = \lim_{n \rightarrow \infty} \widehat{F}(y)$. By lemma Proposition 2, $\lim_{n \rightarrow \infty} \widehat{F}(y) = F_0(y) = \text{Pr}(a_s \leq y)$. Similar arguments as above show that $\sigma_{\gamma,b}^* \xrightarrow{p} \sigma_a$, so that $\text{Pr}(t_b^* \leq y) \rightarrow F_0(y/\sigma_a)$. This establishes that t_b^* has the same asymptotic distribution as \widehat{t} , proving consistency of $y_{\alpha,B}^*$ for y_α . In turn, this consistency result implies that the rejection rule in step 4 yields a test with correct asymptotic Type I error rate.

With one firm, several of the test statistics that HW evaluate follow the above procedure, with one difference. Rather than define R_{sb}^* as in step 3b above, HW use $R_{sb}^* = a_{sb}^*$; equivalently, they use step 3b but impose $\widehat{\beta} = 0$. The distribution of (R_{sb}^*, X_s) thus does not converge to the distribution of (R_s, X_s) as n grows, which might seem problematic. However, the model's independence assumptions show that the distribution (a_{sb}^*, X_s) does still converge to the distribution of (a_s, X_s) , which is sufficient for asymptotically valid inference. The only difference from the argument above in the previous paragraph involves the behavior of β_b^* . Under HW's re-sampling algorithm, it remains true that $\gamma_b^* = R_{n+1,b}^* - X_{n+1} \beta_b^*$. Since it is easy to show that $\text{plim}_{n \rightarrow \infty} \beta_b^* = 0$ under the HW algorithm, γ_b^* has the same asymptotic distribution as $R_{n+1,b}^*$. Since $R_{n+1,b}^* = a_{n+1,b}^*$, the asymptotic distribution of γ_b^* is again $\lim_{n \rightarrow \infty} \widehat{F} = F_0$. The rest of the previous paragraph's argument then follows.

Given the assumed independence of X_s and a_s , the generic residual bootstrap and HW's version of it involve using the empirical distribution of X_s in place of J_{0x} and \widehat{F} in place of F_0 . In other words, each involves estimating the α -quantile of $G_n(\cdot | J_{0x}, F_0)$ with a Monte Carlo estimate of the α -quantile of $G_n(y | \widehat{J}_{0x}, \widehat{F})$. These Monte Carlo estimates satisfy $G_n^*(y | \widehat{J}_{0x}, \widehat{F}) = G_n(y | \widehat{J}_{0x}, \widehat{F}) + o_p(1)$, where the $o_p(1)$ term is Monte Carlo estimation error, which converges to zero in probability.

These Monte Carlo bootstrap procedures are distinct from but closely related to our SQ test. To illustrate this fact, assume for simplicity that we work with standardized fitted abnormal returns, so that $\hat{\sigma}_a = 1$, and $\hat{t} = \hat{\gamma}$. The SQ test involves using the sample α -quantile of the distribution \hat{F} , \hat{y}_α , as the critical value for $\hat{\gamma}$. Because a distribution's quantile function and cdf are inverses, the SQ test essentially involves estimating $G_n(y|J_{0x}F_0)$ with $\hat{F}(y)$. This is asymptotically appropriate, because we have seen that with $\sigma_a = 1$, both \hat{F} and $G_n(y|\hat{J}_{0x}\hat{F})$ converge to F_0 . This discussion shows that both the conventional and HW Monte Carlo residual bootstrap procedures are asymptotically equivalent to our SQ test. Each Monte Carlo procedure chooses a critical value equal to a sample quantile that converges to F_0 . So does our SQ test procedure, but without needing any Monte Carlo sampling.⁴⁵

We conclude that there is no (large-sample) statistical reason to prefer any of these procedures over either of the others. However, our procedure is certainly the simplest to explain, as it involves no Monte Carlo re-sampling.⁴⁶ We believe this is a compelling reason to favor the SQ test in practice. It can be quite difficult to explain Monte Carlo sampling to a lay audience, like policymakers, lawyers or judges. We believe that it will likely be much easier to explain that one need only sort fitted abnormal returns and pick the "right" one as a critical value.

9 Conclusion

In this paper, we have offered a method for inference in event studies with one firm and one event. Single-firm, single-event studies have played an important role in securities litigation. They can also be useful for antitrust policymaking, as well as in many contexts of academic interest. Our method can also be generalized easily to allow for multiple firms and multiple events.

We have shown that our SQ test has a number of good properties. First, it offers asymptotically valid inference, in the sense that its asymptotic Type I error rate equals the desired significance level. Second, it has considerable asymptotic power for empirically relevant abnormal returns distributions. Third, while our SQ test is asymptotically equivalent to various bootstrap procedures, we expect the SQ test will be much easier to explain and justify to a lay audience.

An additional contribution of this paper has been to document the systematic and substantial errors of inference likely to result from inappropriate use of the standard approach to single-firm, single-event studies. The fact of non-normality of abnormal returns distributions, and the basic problems associated with it for event studies have been known and discussed elsewhere, e.g., Hein & Westfall (2004). To our knowledge, though, ours is the first study to document the extent of this problem for event studies using a broad cross-section of firm data. Moreover, we believe that our finding of systematic downward bias in asymptotic Type I error rates is also new. In the context of securities fraud litigation that helps motivate this paper, this evidence suggests

⁴⁵In fact, HW note on page 465 that in the single-firm, single-event case, if the true residuals a_s were known, then "the true bootstrap critical values are evaluated without resorting to Monte Carlo sampling", by using the empirical distribution function of the true residuals, which we called F_n above. The fact that \hat{F} is asymptotically as good an estimate of F_0 as is F_n implies that there is no need for Monte Carlo sampling in practice, either. Moreover, the discussion above shows that the same result holds for the multiple-firm case, and, with a small number of events, for the multiple-events case.

⁴⁶It is also less work from an estimation perspective, though this is a very minor difference given the speed of contemporary computers and the proliferation of canned bootstrap procedures in statistical software.

that over the period 2000-2007, use of the standard approach likely led to pervasive anti-plaintiff evidentiary bias.

Given the fine statistical properties of our SQ test, and the ease with which it can be implemented and explained, we believe it should be adopted in future event studies with small numbers of events or firms.

References

- Andrews, D. (1993), 'Tests for parameter instability and structural change with unknown change point', *Econometrica* .
- Andrews, D. W. K. (2003), 'End-of-sample instability tests', *Econometrica* **71**(6), 1661–94.
- Belsley, D. A., Kuh, E. & Welsch, R. E. (2004), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley-Interscience, New York.
- Bhagat, S. & Romano, R. (2002), 'Event studies and the law: Part i: Technique and corporate litigation', *American Law and Economics Review* **4**, 141–68.
- Bittlingmayer, G. & Hazlett, T. W. (2000), 'Dos kapital: Has antitrust action against microsoft created value in the computer industry?', *Journal of Financial Economics* **55**(3), 329–59.
- Bollerslev, T. & Wooldridge, J. M. (1992), 'Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances', *Econometric Reviews* **11**, 143–72.
- Brown, S. J. & Warner, J. B. (1985), 'Using daily stock returns: The case of event studies', *Journal of Financial Economics* **14**(1), 3–31.
- Campbell, J. Y., Lo, A. W. & MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*, Princeton University Press.
- Carhart, M. M. (1997), 'On persistence in mutual fund performance', *The Journal of Finance* **52**(1), 57–82.
- Chou, P.-H. (2004), 'Bootstrap tests for multivariate event studies', *Review of Quantitative Finance and Accounting* **23**(3), 275–90.
- Chow, G. C. (1960), 'Tests of equality between sets of coefficients in two linear regressions', *Econometrica* **28**(3), 591–605.
- Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**(368), 829–836.
URL: <http://www.jstor.org/stable/2286407>
- Cutler, D. M. & Summers, L. H. (1988), 'The costs of conflict resolution and financial distress: Evidence from the texaco-pennzoil litigation', *The RAND Journal of Economics* **19**(2), 157–72.
URL: <http://www.jstor.org/stable/2555697>
- Davison, A. C. & Hinkley, D. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, United Kingdom.
- Dellavigna, S. & La Ferrara, E. (Forthcoming), 'Detecting illegal arms trade', *American Economic Journal: Applied* .
- Dufour, J.-M. (1980), 'Dummy variables and predictive tests for structural change', *Economics Letters* **6**, 241–47.

- Dufour, J.-M., Ghysels, E. & Hall, A. (1994), ‘Generalized predictive tests and structural change analysis in econometrics’, *International Economic Review* **35**(1).
- Dunbar, F. C. & Mayer, M. K. (2006), *Dura* and the New Vocabulary of Litigation Under Rule 10b-5, Working paper, NERA.
- Efron, B. (1979), ‘Bootstrap methods: Another look at the jackknife’, *The Annals of Statistics* **7**(1).
- Efron, B. & Tibshirani, R. (1994), *An Introduction to the Bootstrap*, Chapman and Hall/CRC.
- Fama, E. (1991), ‘Efficient capital markets ii’, *Journal of Finance* **46**(5), 1575–1617.
- Fama, E. F. & French, K. R. (1992), ‘The cross-section of expected stock returns’, *Journal of Finance* **47**(2), 427–65.
- Fama, E. F. & French, K. R. (1993), ‘Common risk factors in the returns on stocks and bonds’, *Journal of Financial Economics* **33**(1), 3–56.
- Ford, G. S. & Kline, A. D. (2006), Event studies for merger analysis: An evaluation of the effects of non-normality on hypothesis testing,, in ‘Antitrust Policy Issues’, Nova Science Publishers, New York, chapter 8, pp. 135–56.
- Gelbach, J. B., Helland, E. & Klick, J. (2009), Does *Dura* matter? Loss causation and the implications of *Dura Pharmaceuticals v. Broudo*, Working paper, University of Arizona Department of Economics. typescript.
- Greenstone, M., Oyer, P. & Vissing-Jorgensen, A. (2006), ‘Mandated disclosure, stock returns, and the 1964 securities act amendments’, *Quarterly Journal of Economics* **121**(2), 399–460.
- Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.
- Hein, S. E. & Westfall, P. (2004), ‘Improving tests of abnormal returns by bootstrapping the multivariate regression model with event parameters’, *Journal of Financial Econometrics* **2**(3), 451–71.
- Hein, S. E., Westfall, P. & Zhang, Z. (2001), Improvements on event study tests: Bootstrapping the multivariate regression model, Working paper, Texas Tech University. Working Paper.
- Horowitz, J. (2001), The bootstrap in econometrics, in J. J. Heckman & E. E. Leamer, eds, ‘Handbook of Econometrics, Volume 5’, Elsevier Science, pp. 3159–3228.
- Hosken, D. & Simpson, J. D. (2001), ‘Have supermarket mergers raised prices? an event study analysis’, *International Journal of the Economics of Business* **8**(3), 329–42.
- Khotari, S. & Warner, J. B. (n.d.), Econometrics of event studies, in ‘Handbook of Corporate Finance: Empirical Corporate Finance, Volume A’, Elsevier/North-Holland, chapter 1, pp. 3–36.
- Klick, J. & Sitkoff, R. (2008), ‘Agency costs, charitable trusts, and corporate control: Evidence from hershey’s kiss-off’, *Columbia Law Review* **108**(4).

- Kramer, L. A. (2001), Alternative methods for robust analysis in event study applications, *in* ‘Advances in Investment Analysis and Portfolio Management’, Vol. 8, Elsevier Science Ltd., pp. 109–32.
- Li, H., Pincus, M. & Rego, S. O. (2008), ‘Market reaction to events surrounding the sarbanes-oxley act of 2002 and earnings management’, *Journal of Law and Economics* **51**(1), 111–34.
- Mood, A. M. (1950), *Introduction to the Theory of Statistics*, McGraw-Hill Book Company, Inc., New York.
- Newey, W. K. & McFadden, D. (1994), Large sample estimation and hypothesis testing, *in* R. F. Engle & D. McFadden, eds, ‘Handbook of Econometrics’, Vol. 4, North Holland, Amsterdam, pp. 2111–2245.
- Rosenbaum, P. R. (2002), *Observational Studies*, 2nd edn, Springer Verlag, New York.
- Simpson, J. D. (2001), ‘Did may company’s acquisition of associated dry goods corporation reduce competition? an event study analysis’, *Review of Industrial Organization* **18**, 351–62.
- Simpson, J. D. & Hosken, D. (1998), Are retailing mergers anticompetitive? an event study analysis, Working Paper 216, FTC Bureau of Economics Working Paper.
- Taber, C. R. & Conley, T. G. (Forthcoming), ‘Inference with dierece in diereces with a small number of policy changes’, *Review of Economics and Statistics* .
- van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge, U.K.
- Weinstein, M. I. (2008), ‘Don’t buy shares without it: Limited liability comes to American Express’, *Journal of Legal Studies* **37**(1), 189–228.
- White, H. (2001), *Asymptotic theory for econometricians*, revised edn, Academic Press, San Diego, CA.

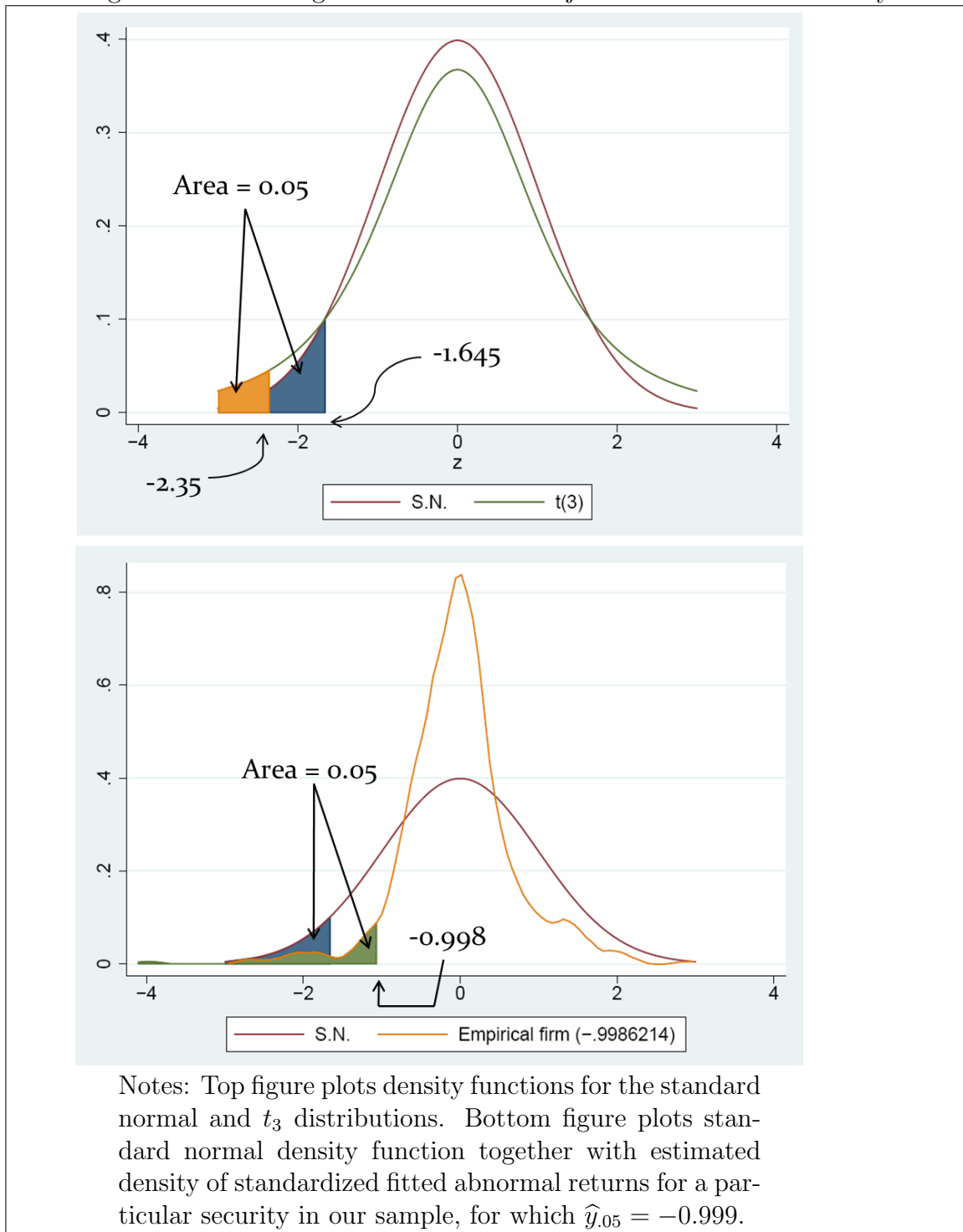
Table 1: List and Description of Variables Used

n	Number of pre-event observations in a generic event study
s, e	Generic date (s), and event date ($e = n + 1$)
R_s^j	Firm j 's return on date s
R_s^m	Market return on date s
X_s	The row vector $(1, R_s^m)$
$\beta^j = (\beta_0^j, \beta_1^j)'$	Usual column vector of coefficients
D_s	Dummy variable indicating whether $s = n + 1 = e$ (the event date)
γ^j	Event effect for firm j
n_j	Number of pre-event observations on firm j in our full sample, S_j
A_s^j	Firm j 's composite abnormal return, including any event effect: $A_s^j = D_s^j \gamma + a_s^j$, for $s = 1, 2, \dots, n_j + 1$
a_s^j	Abnormal return unrelated to event effect, $s = 1, 2, \dots, n_j + 1$
\hat{a}_s^j	Fitted abnormal return: OLS estimate of a_s^j
\tilde{a}_s^j	Standardized fitted abnormal return: $\hat{a}_s^j / \hat{\sigma}_a^j$
F_0^j	True distribution of a_s^j
\hat{F}^j	EDF of \hat{a}_s^j in a sample of size n
$\hat{F}_{n_j}^j$	EDF of \hat{a}_s^j in full sample of n_j observations, i.e., S_j
S_j	Sample of all n_j observations on \hat{a}_s^j : $S_j = \{\hat{a}_s^j\}_{s=1}^{n_j}$
$\hat{\beta}^j, \hat{\gamma}^j$	OLS estimates of β^j and γ^j
$\sigma_a^j, \hat{\sigma}_a^j$	Standard deviation of a_s^j and usual estimate based on OLS market model estimation
$\sigma_\gamma^j, \hat{\sigma}_\gamma^j$	True standard error of γ^j and usual estimate based on OLS market model estimation
α	Desired (nominal) significance level
z_α	α -quantile of standard normal distribution
y_α^j	α -quantile of F_0^j
\hat{y}_α^j	Sample α -quantile of \hat{F}^j
$\lceil x \rceil$	Integer c such that $x - 1 < c \leq x$.
\bar{r}_j	Monte Carlo rejection rate for firm j given desired α
$c(\alpha, n)$	Index of order statistic corresponding to sample α -quantile from sample of size n : $c(\alpha, n) = \lceil \alpha \times (n + 1) \rceil$.

Table 2: Asymptotic Power Under Normality and Empirical Abnormal Returns Distributions

	Value of α		
	0.025	0.05	0.10
$\gamma = -0.5$			
Theoretical, under normality of F_0	0.072	0.126	0.217
Empirical, standard approach	0.050	0.081	0.139
Empirical, SQ Test	0.059	0.116	0.235
$\gamma = -1$			
Theoretical, under normality of F_0	0.169	0.260	0.389
Empirical, standard approach	0.109	0.177	0.330
Empirical, SQ Test	0.126	0.271	0.518

Figure 1: Illustrating Over- and Under-Rejection With Non-Normality



Notes: Top figure plots density functions for the standard normal and t_3 distributions. Bottom figure plots standard normal density function together with estimated density of standardized fitted abnormal returns for a particular security in our sample, for which $\hat{y}_{.05} = -0.999$.

Figure 2: The Cross-Firm Distribution of Actual Versus Permutation-Based Sample .05-Quantiles for Fitted Abnormal Returns

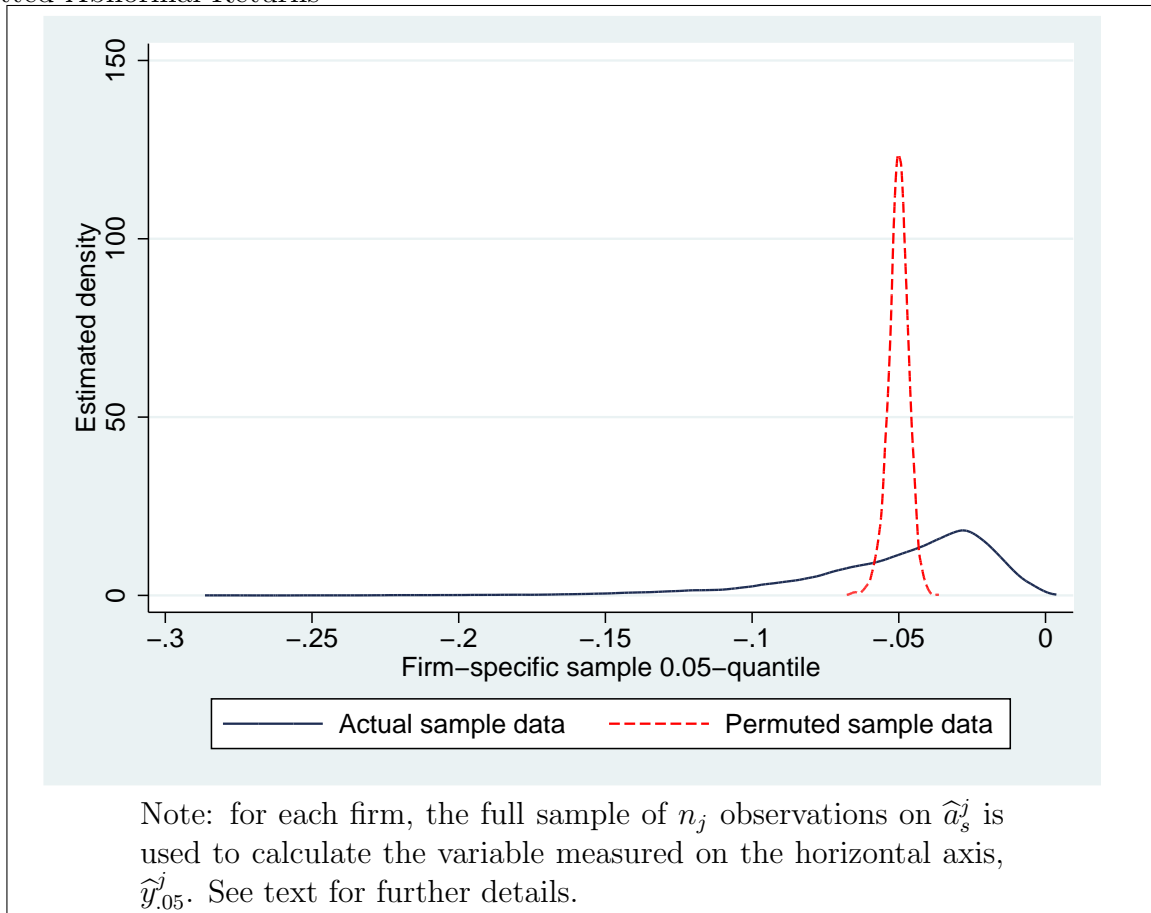


Figure 3: Asymptotic Type I Error Rates for the Standard Approach at Levels $\alpha \in \{0.025, 0.05, 0.10\}$

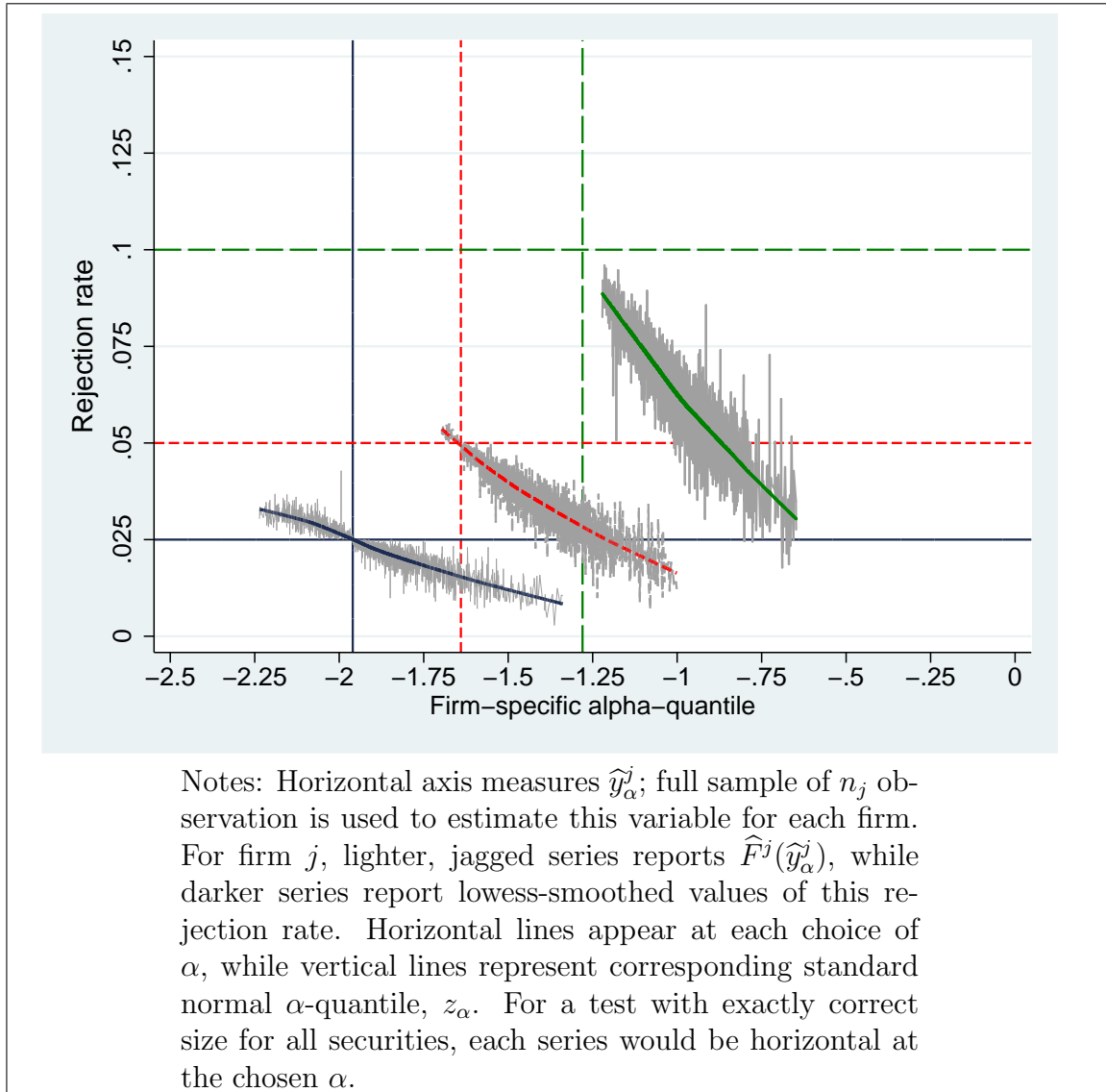


Figure 4: Monte Carlo Size Results for the SQ Test, $\alpha \in \{0.025, 0.05, 0.10\}$

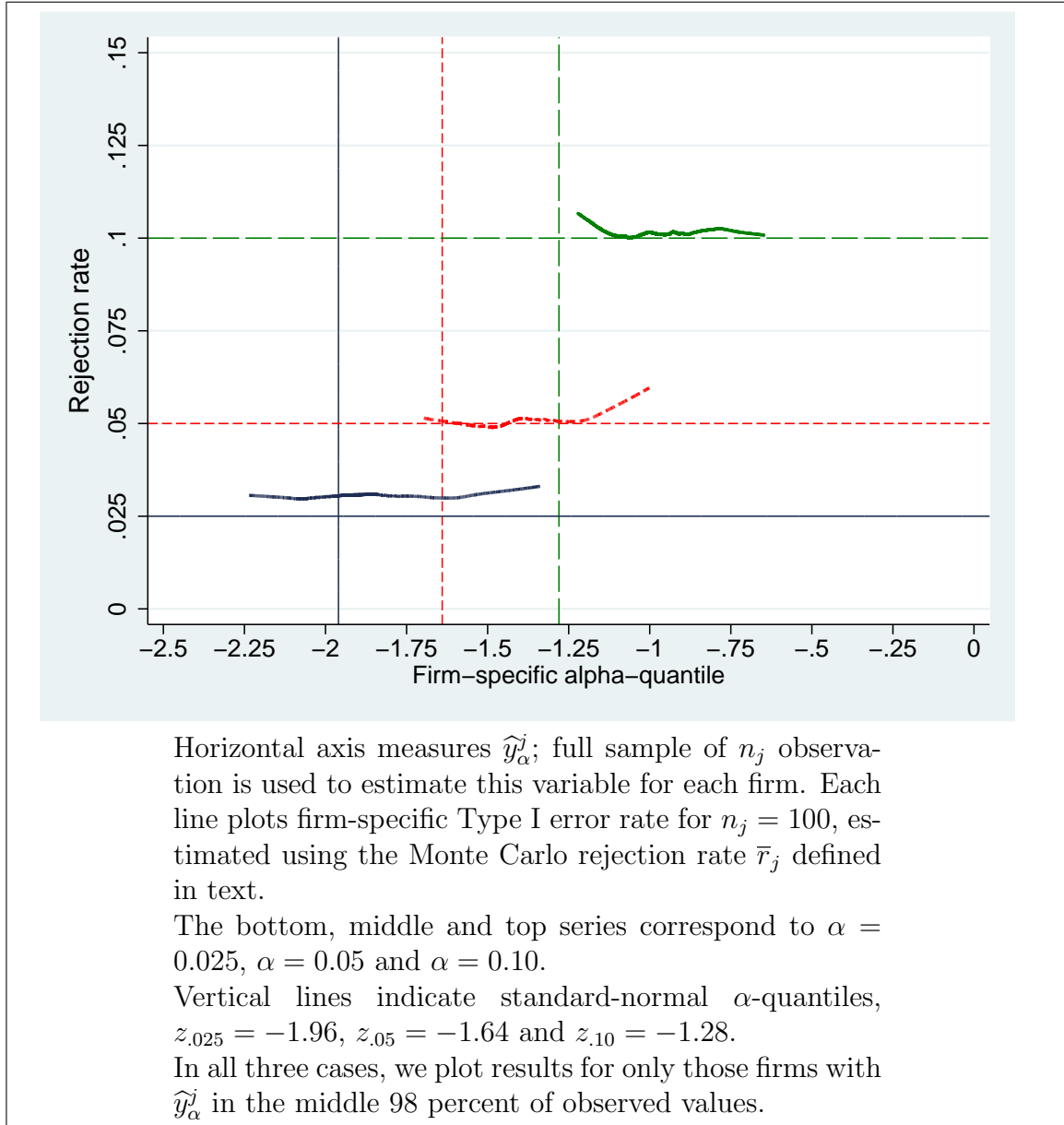
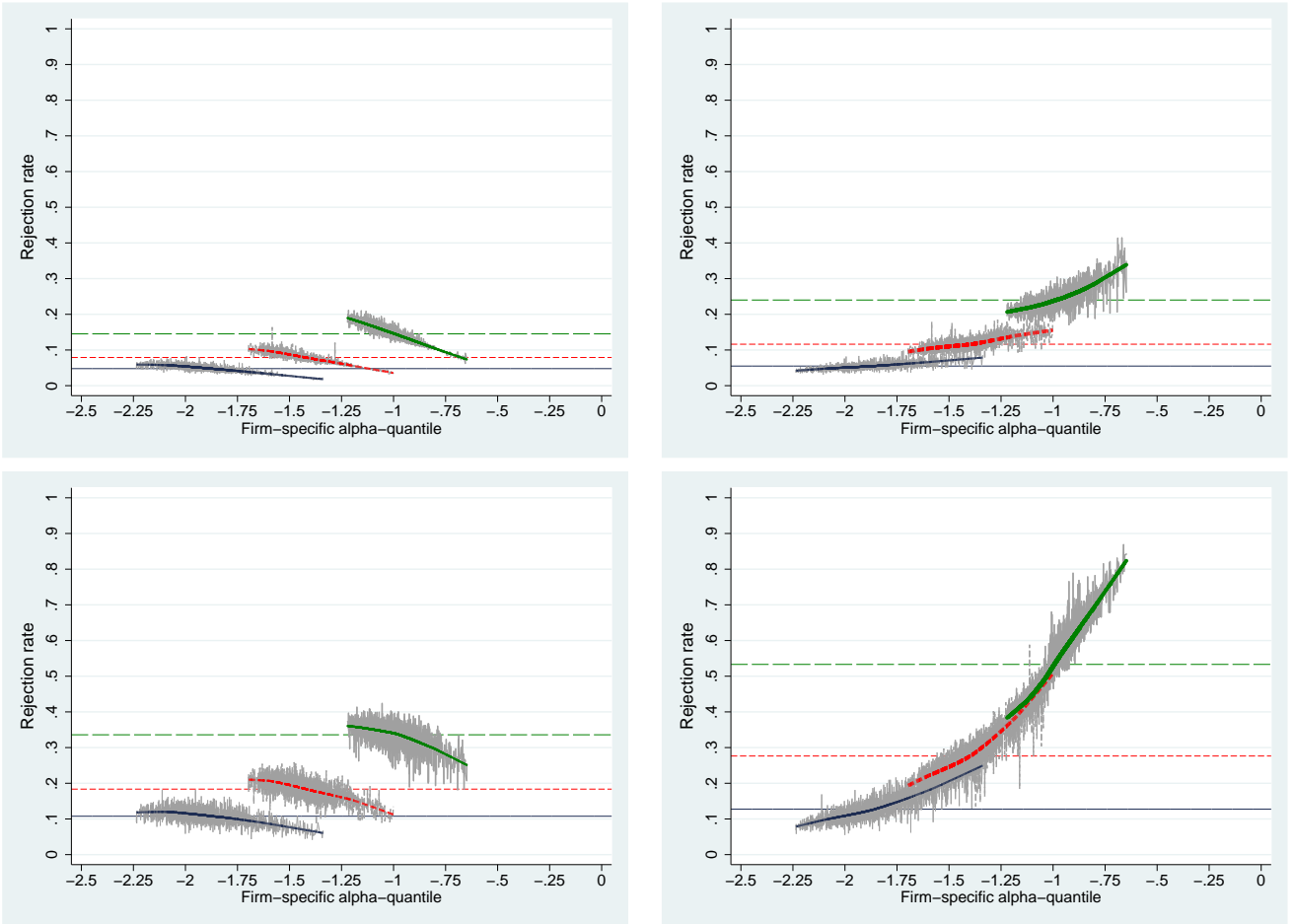


Figure 5: Asymptotic Power for the Standard Approach and the SQ Test, $\gamma \in \{-0.5, -1\}$, $\alpha \in \{0.025, 0.05, 0.10\}$.



Note: Top two graphs have $\gamma = -0.5$, bottom two have $\gamma = -1$. Graphs on left are for standard approach, those on right are for SQ test. Horizontal-axis variable in each graph is $\hat{y}_{0.5}^j$, which we computed using the full sample of n_j observations on each firm j .