Empirical Economics (2024) 66:2435–2450
https://doi.org/10.1007/s00181-023-02530-7

# Single-firm inference in event studies via the permutation test

**Phuong Anh Nguyen**[1,2] · **Michael Wolf**[3]

Received: 27 June 2023 / Accepted: 2 November 2023 / Published online: 23 November 2023
© The Author(s) 2023

**Abstract**
Return event studies generally involve several firms but there are also cases when only one firm is involved. This makes the relevant testing problems, abnormal return and cumulative abnormal return, more difficult since one cannot exploit the multitude of firms (by using a relevant central limit theorem, say) to design hypothesis tests. We propose a permutation test which is of nonparametric nature and more generally valid than the tests that have previously been proposed in the literature in this context. We address the question of the power of the test via a brief simulation study and also illustrate the method with two applications to real data.

**Keywords** Cumulative abnormal return · Event study · Permutation test

**Mathematics Subject Classification** C12 · G14

## 1 Introduction

Return event studies have many applications in accounting and finance; for example, see Campbell et al. (1997, Chapter 4), MacKinlay (1997), Kothari and Warner (2007), Kliger and Gurevich (2014), and the references therein. Given the (intended) brevity

✉ Michael Wolf
michael.wolf@econ.uzh.ch

Phuong Anh Nguyen
npanh@hcmiu.edu.vn

1 Department of Finance and Banking, School of Economics Finance and Accounting, International University, Ho Chi Minh City, Vietnam

2 Vietnam National University, Ho Chi Minh City, Vietnam

3 Department of Economics, University of Zurich, Zürichbergstrasse 14, 8032 Zurich, Switzerland

of this paper, we assume that readers have basic familiarity with the field; otherwise they should feel free to consult the listed references first.[1]

Return event studies are concerned with the question of whether abnormal returns on an event date or, more generally, during a window around an event date (called the event window) are unusually large (in magnitude). To answer this question one carries out a formal hypothesis test where the null hypothesis specifies that the expected value of a certain random variable is zero; if the null hypothesis is rejected, one concludes that the event had an 'impact'.

If there is only one firm under study, the random variable is the abnormal return on the event day itself (AR) or the cumulative abnormal return during the event window (CAR). If there are several firms under study, the respective quantities are averaged across firms. Therefore, the random variable is the average abnormal return on the respective event day (AAR)[2] or the average cumulative abnormal return during the respective event window, which can alternatively be expressed as the cumulative average abnormal return (CAAR).

In most applications there are several firms under study and hence the interest is in testing AAR or CAAR. In such applications one then can exploit the multitude of firms, whose number is generally regarded as the relevant 'sample size', to derive the (approximate) sampling distribution of the chosen test statistic under the null. If the number of firms is large, one can appeal to a suitable central limit theorem and carry out a parametric test.[3] If the number of firms is small, one can carry out a nonparametric test. Again, the reader is referred to the references given above for details.

However, there also exist applications when only one firm is under study, in which case one is interested in testing AR or CAR. This case is more difficult to handle. The classic $t$-test approach is based on the assumption that abnormal returns follow a normal distribution, which is typically violated for financial returns. Gelbach et al. (2013) propose a method for testing AR only that does not require a normal distribution but sill has certain drawbacks; see Sect. 4.3. In this paper, we suggest a permutation test. Such a test does not need to specify a parametric family for the abnormal returns, such as the normal family, and it is valid for testing CAR with a short event window, which includes testing AR as a special case (when the event window is of length one).

## 2 Problem formulation

Abnormal returns are computed based on a given expected-return model for which the user has several choices, such as the constant-mean return model, the market model, the CAPM, or a multi-factor model; our methodology is agnostic concerning this choice.

[1] Apart from return event studies there are also trading-volume event studies and volatility event studies, but those are not the topic of this paper.

[2] More generally, the random variable can be the average abnormal return on any specific day in the event window.

[3] The term "parametric test" is a bit of a misnomer in this context, since one does not need to make the assumption that abnormal returns follow a parametric family, such as a normal distribution; arguably, the term stems from the fact that the (approximate) null distribution of the test statistic follows a parametric distribution, such as the standard normal distribution or a $t$-distribution with certain degrees of freedom.

Returns will be indexed in event time using $t$; note that other people use $\tau$ instead of $t$ for indexing purposes. Defining $t = 0$ as the event date, the range $t = T_1 + 1$ to $t = T_2$ represents the event window with length $m{:=}T_2 - T_1$; of course, it needs to hold that $T_1 + 1 \leq 0 \leq T_2$. In order to unify the exposition, the case $T_1 + 1 = 0 = T_2$ is allowed, in which case the event window consists only of the event day and is thus of length $m = 1$. Furthermore, there is an estimation window that ranges from $t = T_0 + 1$ to $t = T_0 + n \leq T_1$. A leading case in the literature is $T_0 + n = T_1$ in which case the estimation window ends just before the event window begins; for example, see MacKinlay (1997, Section 5). However, this is not a condition we want to impose, as there may be good reasons to use a gap between the two windows.

If $AR_t$ denotes the abnormal return on day $t$, then the cumulative abnormal return during the event window is given by

$$CAR{:=} \sum_{t=T_1+1}^{T_2} AR_t \,. \tag{2.1}$$

One then is interested in testing

$$H_0 : \mathbb{E}(CAR) = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}(CAR) > 0 \tag{2.2}$$

or

$$H_0 : \mathbb{E}(CAR) = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}(CAR) < 0 \tag{2.3}$$

or

$$H_0 : \mathbb{E}(CAR) = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}(CAR) \neq 0 \,. \tag{2.4}$$

The first two testing problems are one-sided whereas the last one is two-sided; the choice of the particular testing problem is up to the user.

Testing CAR includes testing AR as a special case, namely when $T_1 + 1 = 0 = T_2$ and thus $m = 1$, Therefore, there is no need to discuss the problem of testing AR separately, as long as one can devise a method for testing CAR with arbitrary window length $m$, which is exactly the point of our paper.

***Remark 2.1*** (Estimation error in expected-return models) In practice, abnormal returns are computed based on an *estimated* expected-return model, such as the constant-mean model, the market model, or the CAPM. Strictly speaking, the estimation error in the underlying parameter vector of dimension $K$ induces some serial correlation in the abnormal returns. However, as long as the estimation-window size $n$ is not small, this serial correlation is negligible for all practical purposes and can be ignored for the theoretical considerations. As a rule of thumb, $n - K$ should be larger than 100, which is pretty much always the case in practice.

As an example, consider independent and identically distributed (i.i.d.) data $\{x_t\}_{t=1}^n$ and compute abnormal returns based on the constant-mean model ($K = 1$), so that $AR_t{:=}x_t - \bar{x}$ with $\bar{x}{:=}n^{-1} \sum_{t=1}^n x_t$. Then the correlation between $AR_t$ and $AR_j$ is given by $1/(n-1)$ for all $t \neq j$. □

## 3 Classic tests

The classic test statistic for testing problems (2.2)–(2.4) is

$$t_{CAR} := \frac{CAR}{\sqrt{m} s_n} , \tag{3.1}$$

where $s_n^2$ denotes the (unbiased version of the) sample variance of the abnormal returns during the estimation window, that is,

$$s_n^2 := \frac{1}{n-K} \sum_{t=T_0+1}^{T_0+n} AR_t^2 .$$

Here, $K$ denotes the number of parameters that were estimated to compute abnormal returns, which depends on the corresponding model used to that end. For example, $K = 1$ for the constant-mean model, $K = 2$ for the market model and the CAPM, and $K = 4$ for the three-factor Fama–French model (which also includes an intercept). To compute a $p$-value, whose 'formula' depends on which of the testing problems (2.2)–(2.4) has been chosen, one needs to know the (approximate) distribution of the test statistic $t_{CAR}$ under the null. We now detail previous suggestions in the literature, along with the corresponding assumptions to ensure that the test is valid.

MacKinlay (1997) proposes the following approximation (under the null):

$$t_{CAR} \dot{\sim} N(0, 1) , \tag{3.2}$$

that is, the standard normal approximation. This approximation is valid under the assumptions that (i) the data $\{AR_{T_0+1}, \ldots, AR_{T_0+n}, AR_{T_1+1}, \ldots, AR_{T_2}\}$ are independent and identically distributed (i.i.d.) according to a distribution with mean zero and (unknown) variance $\sigma^2 > 0$ and (ii) both $n$ and $m$ are tending to infinity. Concerning (ii), on the one hand, $n$ needs to tend to infinity for $s_n^2 \approx \sigma^2$; on the other hand, $m$ needs to tend to infinity for the standard Lindeberg-Levy central limit to deliver a good approximation. Clearly, (ii) is violated for small event-window sizes, say $m < 15$, which is the leading case of interest in this paper.

Campbell et al. (1997, Section 4.4.3) propose the following approximation (under the null):

$$t_{CAR} \dot{\sim} t_{n-K} , \tag{3.3}$$

which is based on assumption (i) above but strengthened by the additional requirement that the common distribution be a normal one, that is, $N(0, \sigma^2)$. In return, assumption (ii) is no longer required, so that the test can be used also for small event-window sizes $m$. However, it is a well-established fact that daily stock returns generally are not normal because of skewness (that is, asymmetric distribution) and excess kurtosis (that is, heavy-tailed distribution); therefore, using a test based on approximation (3.3) is not safe for small event-window sizes $m$.

We note in passing that as long as $n - K \geq 100$ it does not make a meaningful difference in practice whether one uses approximation (3.2) or (3.3), since the $t$-distribution converges to the standard normal distribution as the degrees of freedom tend to infinity.

***Remark 3.1*** (Safety of a test) To avoid any potential confusion, we now briefly explain what is meant by a test not being safe to use. A hypothesis test has two important features: (i) the (significance) level of the test and (ii) the power of the test. The (significance) level of the test is the probability to reject $H_0$ when it is true (also called the null-rejection probability). Ideally, the actual level should be equal to the nominal level $\alpha$ chosen by the user, the most common choice being $\alpha = 0.05$. In practice, the actual level in finite samples can differ from the nominal level $\alpha$. If the actual level exceeds $\alpha$, the test is invalid and should not be used; such a test is called "liberal". If the actual level is below $\alpha$, then the test is still valid but results in an unnecessary loss of power; such a test is called "conservative". Here, the power of a test is the probability to reject $H_0$ if it is false (or, equivalently, if $H_1$ is true). Obviously, the larger the power, the better. But one should only compare valid tests in terms of their power; it would be pointless to compare the power of a valid test to that of an invalid test, since the invalid test should not be used to begin with. Therefore, if it is not clear whether a test might be invalid, it is not safe to use it. But this is the case when relying on approximation (3.2) or approximation (3.3) for small event-window lengths $m$: If the abnormal returns are not normally distributed, the resulting tests can be liberal or conservative depending on the true distribution, which is unknown. □

Gelbach et al. (2013) study the finite-sample properties of test (3.2) for testing AR (that is, for testing CAR with $m = 1$) based on real-life stock-return data of 3,050 firms contained in the CRSP data base. They find that the test is often conservative, which results in a loss of power.[4]

Alternatively, we can provide a quick 'theoretical' investigation of this issue. It is a well-established fact that (daily) stock returns of many firms have heavy tails. Abstracting from any potential skewness, let us thus assume that, for a given firm, the return distribution is $t_d$, where $d$ denotes the degrees of freedom. Further assume that based on approximation (3.2) ones carries out a two-sided test for AR (that is, for CAR with $m = 1$) at nominal significance level $\alpha$. The test rejects $H_0$ if

$$|t_{CAR}| \geq z_{1-\alpha/2} \,,$$

where $z_\lambda$ denotes the $\lambda$ quantile of $N(0, 1)$. For $n$ tending to infinity, the asymptotic null rejection probability (NRP) is given by

$$\text{Prob}\left\{\left|\sqrt{\frac{d-2}{d}}X\right| \geq z_{1-\alpha/2}\right\} = \text{Prob}\left\{|X| \geq \sqrt{\frac{d}{d-2}} \cdot z_{1-\alpha/2}\right\} \quad \text{where} \quad X \sim t_d \,. \tag{3.4}$$

This is because, for $n$ tending to infinity, $s_n$ converges in probability to $\sqrt{d/(d-2)}$, which is the standard deviation of the $t_d$ distribution. Table 1 presents such asymptotic

[4] Note that they consider a one-sided test of the type (2.3).

**Table 1** Asymptotic null rejection probability (3.4) of the $t$-test for various values of $\alpha$ and $d$. Ideally, the probability should be equal to $\alpha$ always

| $\alpha$ | d = 3 | d = 6 | d = 9 | d = $\infty$ |
|---|---|---|---|---|
| 0.01 | 0.021 | 0.020 | 0.017 | 0.01 |
| 0.05 | 0.043 | 0.533 | 0.534 | 0.05 |
| 0.10 | 0.066 | 0.091 | 0.095 | 0.10 |

NRPs for various values of $\alpha$ and $d$. It is seen that, depending on the scenario, the test can be liberal or conservative.

The question then becomes: Is there a test that is safe to use (or 'valid') under assumption (i) alone? So, on the one hand, we do not want to require the distribution of the abnormal returns to be normal and, on the other hand, we would like the test to be safe to use also for small event-window sizes $m$, including the extreme case $m = 1$. Fortunately, the answer is 'yes'.

## 4 Permutation test

### 4.1 The proposed test

The nonparametric testing method we suggest is not new, but is not very well known among applied researchers (or at least not as well known as it should be) and we have not seen it being promoted or used in this particular context before. The name of the method is *permutation test*.

Under assumption (i), the joint distribution of the data is invariant to permutation (or reordering) of the observations. The combined sample size is $n + m$. Let $X_t := AR_{T_0+t}$ for $t = 1, \ldots, n$ and $X_t = AR_{T_1+t-n}$ for $t = n+1, \ldots, n+m$, so that

$$\{X_1, \ldots X_n, X_{n+1}, \ldots, X_{n+m}\} = \{AR_{T_0+1}, \ldots, AR_{T_0+n}, AR_{T_1+1}, \ldots, AR_{T_2}\} .$$

Next, let $r := \{r_1, \ldots, r_{n+m}\}$ be a permutation (or re-ordering of) of the set of integers $\{1, \ldots, n + m\}$. Note that $(n + m)!$ distinct such permutations exist, where for an integer $d$,

$$d! := d \cdot (d - 1) \cdot \ldots \cdot 2 \cdot 1 .$$

(In words, one says "$d$ factorial".) As an example, there are $3! = 6$ distinct permutations of the set $\{1, 2, 3\}$, given by

$$\{1, 2, 3\}, \{1, 3, 2\}, \{2, 1, 3\}, \{2, 3, 1\}, \{3, 1, 2\}, \{3, 2, 1\} .$$

(Note that the original ordering counts as one of the possible permutations.)

For a given permutation $r$, the corresponding permutation of the $\{X_i\}$ is then implied as $X_i^* := X_{r_i}$ which in return defines the corresponding permutation of the abnormal returns as $AR_{T_0+t}^* := X_t^*$, for $t = 1, \ldots, n$, and $AR_{T_1+t-n}^* := X_t^*$, for $t = n+1, \ldots, n+m$. The point is that under assumption (i) the joint distribution of the permuted abnormal

return is the same as the joint distribution of the original abnormal returns: i.i.d. according to a distribution with mean zero and (unknown) variance $\sigma^2 > 0$.

In a nutshell, the permutation test, in its 'ideal' version, then works as follows. First, set up the test statistic $T$ in a way such that large values 'indicate' the alternative hypothesis, that is,

$$\begin{aligned} T &:= t_{CAR} \quad \text{for testing problem (2.2)}\,, \\ T &:= -\,t_{CAR} \quad \text{for testing problem (2.3)}\,, \text{ and} \\ T &:= |t_{CAR}| \quad \text{for testing problem (2.4)}\,. \end{aligned}$$

Second, for any permutation $r$, denote the value of the test statistic computed from the permuted data $\{AR^*_{T_0+1}, \ldots, AR^*_{T_0+n}, AR^*_{T_1+1}, \ldots, AR^*_{T_2}\}$ by $T^*_r$. Third, compute the $p$-value as

$$\hat{p} := \frac{\#\{T^*_r \geq T\}}{(n+m)!}\,; \tag{4.1}$$

that is, the $p$-value is given by the fraction of test statistics (stemming from all distinct permutations of the data) that are as large or larger than the value of the test statistic computed from the observed data. This algorithm is called 'ideal', since the $p$-value according to formula (4.1) cannot be computed in practice unless the combined sample size $n+m$ is very small, which is not the case in our intended applications; for example, for $n+m = 100$, one obtains $(n+m)! = 100! \approx 9.33 \cdot 10^{157}$.

Therefore, a 'feasible' $p$-value is based on manageable number $B$ of permutations that are selected in a suitable way from universe of all $(n+m)!$ distinct permutations. The 'feasible' $p$-value is then computed as

$$\hat{p} := \frac{\#\{T^*_r \geq T\}}{B}\,;$$

In doing so, it is customary to make the 'identity permutation' one of the selected $B$ permutations, for which then $T^*_r = T$, and draw the remaining $B-1$ permutations at random from the universe of all distinct permutations. In this case, the smallest possible $p$-value is $1/B$, namely if all the test statistics $T^*_r$ based on the $B-1$ randomly drawn permutations are smaller than $T$. It is recommended to choose $B$ as large a possible in practice, depending on one's computational power, but at least $B \geq 10,000$.

Last but not least, how does one draw a permutation of the numbers $\{1, \ldots, n+m\}$ at random? Of course, the exact command depends on one's software but the key term is "drawing without replacement" instead of "drawing with replacement". The mental image is that there is an urn with balls labeled from 1 to $n+m$. Then one draws one ball at a time, at random, without replacement, which results in a random permutation. If one draws with replacement instead, in general some numbers will appear more than once whereas other numbers will not appear at all, and so the resulting sequence is not a permutation.

For completeness, we can now 'summarize' the permutation-test method of constructing a $p$-value by means of the following algorithm.

**Algorithm 4.1** 1. Choose the test statistic $T$ according to the testing problem of interest, (2.2), (2.3), or (2.4), as described just above (4.1).
2. Set $T^*_{r_1} := T$.
3. For $b = 2, \ldots, B$, draw a permutation $r_b$ of the numbers $\{1, \ldots, n+m\}$ at random, permute the data accordingly, and denote the value of the test statistic computed from the permuted data by $T^*_{r_b}$.
4. Compute the $p$-value as

$$\hat{p} := \frac{\#\{T^*_{r_b} \geq T\}}{B} \,. \tag{4.2}$$

By the general results on permutation testing of Lehmann and Romano (2022, Section 17.2.1), the resulting $p$-value (4.2) is exact (or 'perfect') in finite samples; that is, for any $0 < \alpha < 1$,

$$\text{Prob}\big(\hat{p} \leq \alpha\big) = \alpha$$

under assumption (i), the data $\{AR_{T_0+1}, \ldots, AR_{T_0+n}, AR_{T_1+1}, \ldots, AR_{T_2}\}$ are i.i.d. according to a distribution with mean zero and (unknown) variance $\sigma^2 > 0$,

In a sense, the permutation test uses a 'data-based' null distribution to derive the $p$-value, namely the empirical distribution of the $B$ test statistics $\{T^*_{r_1}, \ldots, T^*_{r_B}\}$ computed from permuted data; this is a discrete, nonparametric distribution. On the other hand, the $t$-test uses the $t_{n-K}$ distribution as the null distribution to derive the $p$-value; this is a continuous, parametric distribution. Whereas the former null distribution is always valid by the result stated in the previous paragraph, the latter null distribution is only valid when the abnormal returns follow a normal distribution, which is generally not the case in practice.

It might be instructive to study how the two respective null distributions behave asymptotically, as the size of the estimation window, $n$, tends to infinity whereas the size of the event window, $m$, remains fixed. To this end let $X_1, X_2, \ldots, X_m$ be random variables that are i.i.d. according to the distribution with mean zero and variance $\sigma^2$ that all abnormal returns follow under the null. Then the asymptotic null distribution for the permutation test, if in addition the number of permutations, $B$, tends to infinity, is given by the distribution of the random variable

$$\frac{\sum_{i=1}^{m} X_i}{\sqrt{m}\sigma} \,,$$

which is a distribution with mean zero and variance one.[5] On the other hand, the asymptotic null distribution for the $t$-test is given by $N(0, 1)$, that is, by the standard normal distribution. Therefore, both asymptotic null distributions have mean zero and variance one but only the former is always valid; the latter, again, is only valid when the abnormal returns follow a normal distribution, which is generally not the case in practice.

[5] Relative to the standard normal distribution, the tails of this distribution can be heavier or less heavy; it all depends on the distribution of $X_1$.

### 4.2 Previous uses of permutation tests in event studies

Permutation tests have been proposed in the event-study literature before, albeit in different contexts only (to the best of our knowledge).

Loipersberger (2018) studies, among other things, whether group-aggregate measures such as CAR are 'significantly' different in two groups; here the groups are 'created' by splitting the universe of firms into two groups according to whether a certain measure for a firm (such as "supervision" or "CPI") is weakly greater or strictly smaller than the median number in the entire universe. The test statistic is given by the difference between the CAR values, say, in the two groups. Then one builds up a 'null distribution' by permuting the vector of group memberships across firms and then computing the test statistic for each such permutation.

Bugni et al. (2023) develop a general methodology for testing whether two subsamples have significantly different distributions, with an application to testing for discontinuities in event studies. Their test statistic is given by the Cramèr-von Mises statistic applied to the two empirical distribution functions (computed from each of the two subsamples). Then one builds up a 'null distribution' by permuting the vector of subsample memberships across the entire sample and then computing the test statistic for each such permutation.

Both of these uses of the permutation method address a two-sample testing problem, which is a fundamentally different setting from the one considered in this paper, and therefore they cannot be applied in our context.

### 4.3 Comparison with an alternative proposal for testing AR

Gelbach et al. (2013) convincingly argue that the classic test based on approximation (3.3) is not safe to use in practice for testing AR (that is, for testing CAR with $m = 1$), since it relies on the unrealistic assumption of the abnormal returns having a normal distribution.

As an alternative they propose the sample quantile (SQ) test, which is of nonparametric nature. This test uses sample quantiles of the empirical distribution of the abnormal returns during the estimation window, $\{AR_t\}_{t=T_0+1}^{T_o+n}$, to construct critical values of the test. Gelbach et al. (2013) only provide a description for the one-sided testing problem (2.3), so we take the liberty of providing a 'unified' description for all three testing problems (2.2)–(2.4).

For testing problem (2.2), let $w_t := AR_{t-T_0}$, for $t = 1, \ldots, n$, and let $w := AR_0$; for testing problem (2.3), let $w_t := -AR_{t-T_0}$, for $t = 1, \ldots, n$, and let $w := -AR_0$; for testing problem (2.4), let $w_t := |AR_{t-T_0}|$, for $t = 1, \ldots, n$, and let $w := |AR_0|$.

For $\alpha \in (0, 1)$, and with $\lceil \cdot \rceil$ denoting the 'ceiling operator', let $d_\alpha := \lceil \alpha \cdot n \rceil$ and let $w_{1-\alpha} := w_{(n-d_\alpha+1)}$; hence, $w_{1-\alpha}$ is given by the $(n - d_\alpha + 1)$th value of the order statistics

$$w_{(1)} \leq w_{(2)} \leq \ldots \leq w_{(n)} \, .$$

Then the test rejects $H_0$ at nominal level $\alpha$ if $w > w_{1-\alpha}$.

Compared with the permutation test, the SQ test has the following four disadvantages:

D-1 The test as proposed by the authors only applies to testing AR, that is, testing CAR with $m = 1$. It might be possible to extend the proposal to testing CAR with $m > 1$, but this is not discussed in the paper.

D-2 The test is 'only' asymptotically valid, that is, the null rejection probability (NPR) can exceed $\alpha$ in finite samples, and is guaranteed to be bounded above by $\alpha$ only in the limit as $n$ tends to infinity; the fact that the finite-sample NPR can be above $\alpha$ is evident from the Monte Carlo study in Gelbach et al. (2013, Section 5.2).

D-3 The authors stress the importance of choosing $n$ such that $\alpha \cdot n$ is an integer; otherwise the finite-sample performance "will yield upward distortions in actual size". Depending on data availability, this means that some observations might have to be discarded, which would result into a loss of information.

D-4 The test can only be carried out for given nominal significance level $\alpha$; it is not possible to compute a $p$-value (directly). In principle, one could compute a $p$-value indirectly as the smallest significance level $\alpha$ at which $H_0$ can 'just' be rejected. But there are two problems. First, doing so is clearly cumbersome and involves a 'trial-and-error' approach. Second, and more importantly, for this indirect method to yield a valid $p$-value, the test must be valid for any $\alpha \in (0, 1)$; but by D-3, the test "will yield upward distortions in actual size" if $\alpha \cdot n$ is not an integer. Therefore, even the indirect approach cannot be used to compute a 'precise' $p$-value, meaning a $p$-value that has (at least) three significant digits.

### 4.4 Nonrobustness to event-induced increase in variance

We deem it prudent to point out that all three tests that have been discussed — $t$-test, permutation test, and SQ test — are not robust to event-induced increase in variance.

To illustrate, consider as an example an event window of a single day (that is, $m = 1$) in which case the two-sided testing problem specifies to

$$H_0 : \mathbb{E}(AR) = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}(AR) \neq 0 \,.$$

Further, for simplicity, assume that abnormal returns are normally distributed. Our assumption under $H_0$ then specifies to the assumption that the data $\{AR_{T_0+1}, \ldots, AR_{T_0+n}, AR_{T_1+1}\}$ are i.i.d. according to $N(0, \sigma^2)$. If instead $AR_{T_1+1}$ is distributed according to $N(0, \tilde{\sigma}^2)$ with $\tilde{\sigma}^2 > \sigma^2$, the probability of rejecting $H_0$ is not controlled at the nominal level $\alpha$. In fact, the rejection probability of all three tests tends to one as $\tilde{\sigma}^2$ tends to infinity.

Therefore, to be allowed to interpret a rejection of $H_0$ as evidence for $\mathbb{E}(AR) \neq 0$, one must assume that the event, if it has any effect, only changes the mean of $AR_{T_1+1}$ but not its variance, such that $AR_{T_1+1}$ is distributed according to $N(\gamma, \sigma^2)$ with $\gamma \neq 0$.

This reasoning carries over to multi-day event windows (that is, $m > 1$) and abnormal returns that are not normally distributed: Any effect of the event should only 'shift' the distributions of the abnormal returns (up or down) but leave the shape of the distribution otherwise unchanged, and thus in particular leave the variance unchanged.

The nonrobustness to event-induced increase in variance of the three tests is undesirable but also impossible to fix, in the sense that it is not possible to disentangle statistically a change of the mean of the abnormal returns in the event window from a potential change of the variance at the same time, at least when the event window is short and as short as a single day.

### 4.5 Extension to testing CAAR

As stated before, in most event studies there are several firms under study and the interest is in testing CAAR, of which AAR is a special case (when the event window is of size $m = 1$). If the number of firms is 'sufficiently' large, one can use parametric test statistics; as a rule of thumb, $N \geq 30$ firms can be considered sufficient. For a smaller number of firms, one can use nonparametric test statistics; as a rule of thumb, $N \geq 10$ firms can be considered sufficient.[6]

But there might be applications when the number of firms is in the single digits and as small as $N = 2$. In such cases, even nonparametric test statistics are generally not viable. On the other hand, one can extend the permutation test for testing CAR outlined above to such applications. Once one prescribes how to permute the joint data comprising all the firms, the way the test is carried out is similar to testing CAR, and thus the details are left to the reader. In prescribing how to permute the joint data, we shall consider two settings.

In the first setting, there is no overlap between the 'combined' windows (estimation window together with event window) of the various firms. In this setting, one would permute 'independently' with respect to firms; in other words, one would permute the firm-specific data one firm at a time, using independently drawn permutations.

In the second setting, there is a common estimation window together with a common event window for all firms. In this setting, one would always apply the same permutation to all the firms together in order to preserve any (potential) across-firm dependence structure.

## 5 A brief power comparison

Applied researchers might be concerned about whether the permutation test results in a loss of power compared to the $t$-test or, in case of testing AR (that is, testing CAR with $m = 1$) compared to the SQ test. We address this concern via a brief Monte Carlo study.

If assumption (i) is strengthened to: the data $\{AR_{T_0+1}, \ldots, AR_{T_0+n}, AR_{T_1+1}, \ldots, AR_{T_2}\}$ are i.i.d. according to a *normal* distribution with mean zero and (unknown) variance $\sigma^2 > 0$, then both the $t$-test and the permutation test have exact (or 'perfect') level $\alpha$ in finite samples. Therefore, the normal setting is the fair setting to compare power. If instead we chose a setting where the $t$-test is conservative, this would give an unfair advantage to the permutation test; on the other hand, if we chose a setting

[6] Even for $N \geq 30$ firms there might be good reasons to prefer nonparametric test statistics but this issue is not the concern of our paper.

**Table 2** Empirical powers of various tests for four different scenarios when data come from a normal distribution with common variance

| Scenario | $t$-test | Perm test | SQ test |
|---|---|---|---|
| S-1 | 0.51 | 0.50 | 0.50 |
| S-2 | 0.51 | 0.51 | NA |
| S-3 | 0.52 | 0.51 | NA |
| S-4 | 0.51 | 0.50 | NA |

*Perm test* permutation test, *NA* not available

where the $t$-test is liberal, this would give an unfair advantage to the $t$-test; also recall that a liberal test should not be used to begin with.

Therefore, we can make a fair power comparison by considering the following setting: $\{AR_{T_0+1}, \ldots, AR_{T_0+n}\}$ are i.i.d. $\sim N(0, \sigma^2)$ whereas $\{AR_{T_1+1}, \ldots, AR_{T_2}\}$ are independently distributed with $AR_{T_1+j} \sim N(\mu_j, \sigma^2)$. Also denote $\mu := (\mu_1, \ldots, \mu_m)$. We shall consider four scenarios where $H_1$ is true; in all scenarios, $n = 120$ and $\sigma^2 = 1$.

- Scenario 1 (S-1): $m = 1$ with $\mu = 2$
- Scenario 2 (S-2): $m = 5$ with $\mu = 0.9 \cdot (1, 1, 1, 1, 1)$
- Scenario 3 (S-3): $m = 5$ with $\mu = 0.3 \cdot (1, 2, 3, 4, 5)$
- Scenario 4 (S-4): $m = 5$ with $\mu = 1 \cdot (0, 0, 1, 5, 0)$

S-1 corresponds to testing AR and is the only scenario where the SQ test can also be used. S-2 corresponds to a constant effect during the event window. S-3 corresponds to an increasing effect during the event window. S-4 corresponds to a small effect on the event day, a large effect on the day after, and no effects on the other days. For a given scenario, the numerical values in $\mu$ were chosen by trial and error to give empirical powers in the neighborhood of 0.5, which is the most 'discriminating' region to distinguish between various tests with respect to their power.

Empirical powers are computed based on 100,000 Monte Carlo repetitions. The permutation test is based on $B = 1,000$ permutations. The testing problem is (2.4), so the tests are two-sided. Table 2 presents the results. It can be seen that in all four scenarios the permutation test does not lead to a meaningful loss in power, if in any loss at all. Therefore, at least at the conventional significance level $\alpha = 0.05$, applied researchers should not be deterred from using the permutation test because of power concerns.

## 6 Two real-life examples

### 6.1 Pirnik v. Fiat Chrysler Autos

During the legal case "Pirnik v. Fiat Chrysler Autos" decided on 26-June-2018, event-study methodology was used by the plaintiffs to identify abnormal price movements in response to six allegedly corrective disclosures within the context of the "Dieselgate" scandal. If found significant, the price movements would count as evidence for a "price maintenance" scenario under which the stock price was kept inflated by prior

**Table 3** Test statistics and two-sided $p$-values for the null hypothesis $H_0 : \mathbb{E}(CAR) = 0$ for various event-window sizes $m$

| $t_{CAR}$ | −2.158 | −0.851 | −1.930 | −2.469 |
|---|---|---|---|---|
| | m = 1 | m = 3 | m = 5 | m = 7 |
| $t$-test | 0.033 | 0.396 | 0.056 | 0.015 |
| Permutation test | 0.049 | 0.298 | 0.065 | 0.033 |

**Table 4** Test statistics and two-sided $p$-values for the null hypothesis $H_0 : \mathbb{E}(AR) = 0$ for all seven days (individually) during the event window; of course, here $CAR = AR$ always

| $t_{CAR}$ | −1.158 | −1.304 | 1.652 | −2.158 | −0.968 | −1.538 | −1.057 |
|---|---|---|---|---|---|---|---|
| | t = -3 | t = -2 | t = -1 | t = 0 | t = 1 | t = 2 | t = 3 |
| $t$-test | 0.249 | 0.195 | 0.101 | 0.033 | 0.335 | 0.127 | 0.293 |
| Permutation test | 0.166 | 0.148 | 0.091 | 0.049 | 0.217 | 0.091 | 0.190 |
| SQ test | $> 0.1$ | $> 0.1$ | $\leq 0.1$ | $\leq 0.05$ | $> 0.1$ | $\leq 0.1$ | $> 0.1$ |

misinformation. The corrective disclosure event studied hereafter took place on 23-May-2016 in the form of a report by Germany's Bild newspaper which stated that the carmaker could be prohibited from selling cars in Germany if evidence that it had disregarded emissions rules was found. The event was deemed significant at a 0.9927 confidence level, which is equivalent to a 0.0073 significance level; for example, see (Pirnik v. Fiat Chrysler Autos 2018 p. 6).

In our revisiting of the event, we shifted the estimation window back to end four trading days prior to the event, and its size was set to $n = 120$ trading days. For purpose of illustration, we consider event windows of size $m = \{1, 3, 5, 7\}$, always centered at the event day. The abnormal returns are computed using the market model with the S&P 500 index serving as the market proxy. The abnormal returns during the event window are given by $\{-0.0183, -0.0206, 0.0261, -0.0341, -0.0153, -0.0243, -0.0167\}$, so the abnormal return on the event day itself was $-0.0341$.

Table 3 presents the results. For each event-window size $m$, the table lists the value of the test statistic $t_{CAR}$ as well as two-sided $p$-values corresponding to testing problem (2.4) for the $t$-test based on (3.3) (with $K = 2$ for the market model) and for the permutation test (using $B = 100,000$). One can see that the testing methods are in general agreement here: $H_0$ can be rejected at significance level $\alpha = 0.05$ for $m = 1, 7$ and at significance level $\alpha = 0.1$ for $m = 5$.

In litigation cases, often each day during an event window is of (individual) interest as well. Table 4 presents the results. For each day $t \in \{-3, -2, \ldots, 3\}$, the table lists the value of the test statistic $t_{CAR}$ as well as two-sided $p$-values corresponding to testing problem (2.4) for the $t$-test based on (3.3) (with $K = 2$ for the market model) and for the permutation test (using $B = 100,000$); note that here $m = 1$ always. In addition the table presents a '$p$-value range' for the SQ test of Sect. 4.3: "$> 0.1$" means that the test fails to reject at significance level $\alpha = 0.1$; "$\leq 0.1$" means that the test rejects at significance level $\alpha = 0.1$ but fails to reject at significance level $\alpha = 0.05$; "$\leq 0.05$" means that the test rejects at significance level $\alpha = 0.05$. Recall that it is not possible to compute a 'precise' $p$-value for the SQ test; see disadvantage D-4 in

**Table 5** Test statistics and two-sided $p$-values for the null hypothesis $H_0 : \mathbb{E}(CAR) = 0$ for various event-window sizes $m$

| $t_{CAR}$ | −5.551 | −3.953 | −2.760 | −2.122 |
|---|---|---|---|---|
| | m = 1 | m = 3 | m = 5 | m = 7 |
| $t$-test | 0.000 | 0.000 | 0.007 | 0.036 |
| Permutation test | 0.008 | 0.004 | 0.016 | 0.042 |

Sect. 4.3. Also, since $0.01 \cdot 120 = 1.2$ is not an integer, based on the recommendation of Gelbach et al. (2013), we prefer not to carry out the SQ test at significance level $\alpha = 0.01$ to avoid an upward distortion in size; see disadvantage D-3. The following differences can be seen: Based on both the permutation test and the SQ test $H_0$ can be rejected at significance level $\alpha = 0.1$ for $t = -1$ and 2, whereas based on the $t$-test $H_0$ cannot be rejected for either of these days (at significance levels $\alpha = 0.1$ or below).

### 6.2 Twist bioscience corporation securities litigation

Our second example is from an ongoing case. The biotech research firm Twist Bioscience is being accused of false reporting of capital expenditures and cross margins by the activist short seller and research house Scorpion Capital. The event the market had to price was ambiguous: a lengthy research report tainted in its objectivity by the obvious interest of the short seller to downgrade the firm. The report was released on 15-November-2022 and describes Twist Bioscience as a "cash-burning inferno that […] will end in bankruptcy." Although the stock took a punch, it may not have done so at the extent intended — particularly from today's perspective where the stock seems to have stabilized — thus questioning which conclusion could be drawn from the corrective disclosure event.

In our revisiting of the event, we shifted the estimation window back to end four trading days prior to the event, with its size set to $n = 120$ trading days. For purpose of illustration, we consider event windows of size $m = \{1, 3, 5, 7\}$, always centered at the event day. The abnormal returns are computed using the market model with the S&P 500 index serving as the market proxy. The abnormal returns during the event window are given by $\{0.0468, 0.1236, -0.0130, -0.2284, -0.0403, -0.0958, -0.0239\}$, so the abnormal return on the event day itself was $-0.2284$.

Table 5 presents the results. For each event-window size $m$, the table lists the value of the test statistic $t_{CAR}$ as well as two-sided $p$-values corresponding to testing problem (2.4) for the $t$-test based on (3.3) (with $K = 2$ for the market model) and for the permutation test (using $B = 100,000$). One can see that the testing methods are in general agreement here: $H_0$ can be rejected at significance level $\alpha = 0.05$ for all event-window sizes $m$.

Table 6 presents the results for the individual days during the event window. For each day $t \in \{-3, -2, \ldots, 3\}$, the table lists the value of the test statistic $t_{CAR}$ as well as two-sided $p$-values corresponding to testing problem (2.4) for the $t$-test based on (3.3) (with $K = 2$ for the market model) and for the permutation test (using

**Table 6** Test statistics and two-sided $p$-values for the null hypothesis $H_0 : \mathbb{E}(AR) = 0$ for all seven days (individually) during the event window; of course, here $CAR = AR$ always

| $t_{CAR}$ | 1.137 | 3.004 | −0.316 | −5.551 | −0.979 | −2.328 | −0.581 |
|---|---|---|---|---|---|---|---|
| | t = -3 | t = -2 | t = - 1 | t = 0 | t = 1 | t = 2 | t = 3 |
| $t$-test | 0.258 | 0.003 | 0.753 | 0.000 | 0.329 | 0.022 | 0.562 |
| Permutation test | 0.207 | 0.017 | 0.769 | 0.008 | 0.298 | 0.041 | 0.595 |
| SQ test | $> 0.1$ | $\leq 0.05$ | $> 0.1$ | $\leq 0.05$ | $> 0.1$ | $\leq 0.05$ | $> 0.1$ |

$B = 100,000$); note that here $m = 1$ always. In addition the table presents a '$p$-value range' for the SQ test of Sect. 4.3: "$> 0.1$" means that the test fails to reject at significance level $\alpha = 0.1$; "$\leq 0.1$" means that the test rejects at significance level $\alpha = 0.1$ but fails to reject at significance level $\alpha = 0.05$; "$\leq 0.05$" means that the test rejects at significance level $\alpha = 0.05$. Recall that it is not possible to compute a 'precise' $p$-value for the SQ test; see disadvantage D-4 in Sect. 4.3. Also, since $0.01 \cdot 120 = 1.2$ is not an integer, based on the recommendation of Gelbach et al. (2013), we prefer not to carry out the SQ test at significance level $\alpha = 0.01$ to avoid an upward distortion in size; see disadvantage D-3. The following difference can be seen: For $t = -2$, $H_0$ can be rejected at significance level $\alpha = 0.01$ based on the $t$-test but 'only' at significance level $\alpha = 0.05$ based on the permutation test and the SQ test.

***Remark 6.1*** (Safety of a test, revisited) The fact that in most of the testing problems above the two methods that can be applied for testing CAR in general (that is, for $m \geq 1$), the $t$-test and the permutation test, come to the same decision — in terms of rejecting $H_0$ or not at a given significance level $\alpha$ — should not come as a surprise, at least to anyone versed in the field of statistics. If the test statistic is 'close enough' to zero, respectively 'away enough' from zero, then all semi-reasonable testing methods, even if they are not entirely safe to use, will come to the same decision: do not reject $H_0$, respectively reject $H_0$. It is in the relatively rare instances of a 'borderline' test statistic where differences between a safe test and an unsafe test can be observed. The fact that an unsafe test produces the correct decision most of the time is not a justification for using it. To make an analogy: Wearing a seat belt when driving a car most of the time makes no difference compared to not wearing one; still, the prudent thing is to wear one all the time. □

## 7 Conclusion

This paper has proposed the use of the permutation test for single-firm event studies when the event window is short and as short as a single day. Unlike the $t$-test, the permutation test does not rely on the (unrealistic) assumption of the abnormal returns being normally distributed. Unlike the SQ test, which is also of nonparametric nature, the permutation test has a null-rejection probability equal to the nominal level in finite samples already and not only asymptotically (as the size of the estimation window

tends to infinity). A simulation study has demonstrated that a potential loss of power compared to the $t$-test and the SQ test is not a concern in practice. An application to two real-life data sets has shown that the permutation test can lead to more significant findings than the $t$-test and can be more widely applied than the SQ test, which is restricted to event windows of a single day. A possible topic for future research is the use of the permutation test for multiple-firm event studies when the number of firms is small.

**Funding** Open access funding provided by University of Zurich The authors did not receive support from any organization for the submitted work.

**Data availibility** The data that support the findings of this study are available from the corresponding author upon request.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

Bugni FA, Li J, Li Q (2023) Permutation-based tests for discontinuities in event studies. Quant Econ 14(1):37–70

Campbell J, Lo A, MacKinlay C (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton

Gelbach JB, Helland E, Klick J (2013) Valid inference in single-firm, single-event studies. Am Law Econ Rev 15(2):495–541

Kliger D, Gurevich G (2014) Event studies for financial research. Palgrave Macmillan, New York

Kothari S, Warner J (2007) Econometrics of event studies. In: Eckbo BE (ed) Handbook of empirical corporate finance: empirical corporate finance, vol 1. Elsevier, Amsterdam, pp 3–36

Lehmann EL, Romano JP (2022) Testing Statistical Hypotheses, 4th edn. Springer, New York

Loipersberger F (2018) The effect of supranational banking supervision on the financial sector: event study evidence from Europe. J Bank Finance 91:34–48

MacKinlay AC (1997) Event studies in economics and finance. J Econ Lit 35(1):13–39

Pirnik v. Fiat Chrysler Autos (2018) 327 f.r.d. 38 (s.d.n.y. 2018). https://casetext.com/case/pirnik-v-automobiles-4/case-details

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.