University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 425

A Note on Testing AR and CAR for Event Studies

Phuong Anh Nguyen and Michael Wolf

January 2023

A Note on Testing AR and CAR for Event Studies

Phuong Anh Nguyen
Department of Finance and Banking
School of Economics, Finance and Accounting
International University
Vietnam National University
Ho Chi Minh City, Vietnam
npanh@hcmiu.edu.vn

Michael Wolf*
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

January 2023

Abstract

Return event studies generally involve several companies but there are also cases when only one company is involved. This makes the relevant testing problems, abnormal return (AR) and cumulative abnormal return (CAR), more difficult since one cannot exploit the multitude of companies (by using a relevant central limit theorem, say). We propose a permutation test that is valid under weaker conditions than the tests that have previously 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.

KEY WORDS: Cumulative abnormal return, event study, permutation test.

JEL CLASSIFICATION NOS: C12, G14.

*Corresponding author. Mailing address: Zürichbergstrasse 14, CH-8032 Zurich, Switzerland.

1 Introduction

Return event studies have many applications in accounting and finance; e.g., 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 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 company 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 companies under study, the respective quantities are averaged across companies. So 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 companies under study and hence the interest is in testing AAR or CAAR. In such applications one then can exploit the multitude of companies, 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 companies is large, one can appeal to a suitable central limit theorem and carry out a parametric test.[3] If the number of companies 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 company is under study, in which case one is interested in testing AR or CAR. This case is more difficult to handle since, with the approaches so far suggested in the literature, one needs to make the assumption that abnormal returns follow a normal distribution or, at least when testing CAR, that the event window is

[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.

unduly large. Therefore, as an alternative approach that is valid more generally, 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. Returns will be indexed in event time using t; note that other people use τ instead of t. Defining $t = 0$ as the event date, $t = T_1 + 1$ to $t = T_2$ represents the event window with length $m := T_2 - T_1$. We need of course $T_1 + 1 \leq 0 \leq T_2$. 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 one. 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 $\{x_i\}_{i=1}^n$ and compute abnormal returns based on the constant-mean model ($K = 1$), so that $AR_i := x_i - \bar{x}$ with $\bar{x} := n^{-1} \sum_{i=1}^n$. Then the correlation between AR_i and AR_j is given by $1/(n-1)$ for all $i \neq j$. ■

3 Previous Tests

The common 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,

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

where 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 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. There can be no hard-and-fast rule, but in our view one would need something like $n \geq 150$ and $m \geq 30$ in order to rely on approximation (3.2). These conditions are rarely met in practice, foremost because event windows of length $m \geq 30$ are rare. Moreover, if one had an event window of length $m > 30$, then, in our view, a better test is available this day and age: Simply use the (studentized) bootstrap based on the event-window sample $\{AR_{T_1+1}, \ldots, AR_{T_2}\}$ to test for a (common) mean of zero; for example, see Davison and Hinkley (1997, Chapter 4).

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.

As long as $n - K > 100$, which is pretty much always the case in practice, it matters very little whether one use approximation (3.2) or approximation (3.3), since the t-distribution 'converges' to the standard normal distribution as the degrees of freedom tend to infinity. The important point is that the resulting tests can also be used for a small event-window size m if the distribution of the abnormal returns is normal, which is a very strong assumption. Indeed, 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; as a rule of thumb $m < 30$ can be considered not safe. In the end, the limitation is the same as for the test based on approximation (3.2), the reason being that both approximations yield nearly identical results when $n - K > 100$.

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 to important features: (i) the (significance) level of the test and (ii) the power of the test. The level of the test is the probability to reject H_0 if it is true (also called the null-rejection probability). Ideally, the level should be equal to the nominal level α 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 α. If the 'actual' level exceeds α, the test is 'invalid' and should not be used; such a test is called "liberal". If the 'actual' level is below α, then the test is 'valid' but results in an unnecessary loss of power; such a test is called "conservative". 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 the power 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 using approximation (3.2) or approximation (3.3). If the data are not normally distributed, the test can be liberal or conservative depending on the true distribution, which is unknown. ■

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

The 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. 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 $X_i := AR_{T_0+i}$ for $i = 1, \ldots, n$ and $X_i = AR_{T_1+i-n}$ for $i = n + 1, \ldots, n + m$, 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 $AR^*_{T_0+i} := X_i^*$ for $i = 1, \ldots, n$ and $AR^*_{T_1+i-n} := X_i^*$ for $i = n + 1, \ldots, n + m$. 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,

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

Second, for any permutation r, denote the value of the test statistic computed form 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 $\sigma^2 > 0$,

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. 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 α 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 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 $AR_{T_1+j} \sim N(\mu_j, \sigma^2)$; in particular, if $\mu_1 = \ldots = \mu_m = 0$, then H_0 is true. Also denote $\mu := (\mu_1, \ldots, \mu_m)$. We shall consider four scenarios; 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. 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 μ 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 10,000 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. The rule is that H_0 is rejected if the p-value of the test is less than or equal to $\alpha = 0.05$.

Table 5.1 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.

Scenario	t-test	Perm test
S-1	0.51	0.49
S-2	0.51	0.51
S-3	0.52	0.51
S-4	0.51	0.50

Table 5.1: Empirical powers of the t-test and the permutation test (Perm test) for four different scenarios when data come from a normal distribution with common variance.

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 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 6.3 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 $m = 5$.

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

In litigation cases, often each day during an event window is of (individual) interest as well. Table 6.4 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. As before, of course, both tests reject H_0 for $t = 0$ (the event day) at significance level $\alpha = 0.05$. For the remaining days, there are two differences, though. Based on the permutation 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 any of the remaining days (at significance levels $\alpha = 0.1$ or below).

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

Table 6.2: Test statistics and two-sided p-values for the null hypothesis $H_0 : \mathbb{E}(CAR) = 0$ for all seven days (individually) during the event window; of course, here $CAR = AR$ always.

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 company. 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 $(0.04680.1236 - 0.0130 - 0.2284 - 0.0403 - 0.0958 - 0.0239)$, so the abnormal return on the event day itself was -0.2284.

Table 6.3 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 all 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

Table 6.3: Test statistics and two-sided p-values for the null hypothesis $H_0 : \mathbb{E}(CAR) = 0$ for various event-window sizes m.

Table 6.4 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 $B = 100,000$); note that here $m = 1$ always. One can see that, again, the testing methods are in general agreement here: H_0 can be rejected at significance level $\alpha = 0.05$ for $t \in \{-2, 0, 2\}$, and it cannot be rejected for any of the remaining days (at significance levels $\alpha = 0.1$ or below).

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

Table 6.4: Test statistics and two-sided p-values for the null hypothesis $H_0 : \mathbb{E}(CAR) = 0$ for all seven days (individually) during the event window; of course, here $CAR = AR$ always.

Remark 6.1. The fact that in most of the testing problems above the two methods, t-test and permutation test, come to the same decision — in terms of rejecting H_0 or not — 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 to use 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 Extension to Testing CAAR

As stated before, in most event studies there are several companies 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, $N \geq 30$ companies can be considered sufficient. For a smaller number of companies, one can use nonparametric test statistics; as a rule of thumb, $N \geq 10$ companies can be considered sufficient.[4]

But there might be applications when the number of companies 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 companies, 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 companies. In this setting, one would permute 'independently' with respect to companies; in other words, one would permute the company-specific data one company 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 companies. In this setting, one would always apply the same permutation to all the companies together in order to preserve any (potential) across-company dependence structure.

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

References

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

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press, Cambridge.

Kliger, D. and Gurevich, G. (2014). *Event Studies for Financial Research.* Palgrave Macmillan, New York.

Kothari, S. and Warner, J. (2007). Econometrics of event studies. In Eckbo, B. E., editor, *Handbook of Empirical Corporate Finance: Empirical Corporate Finance, Volume 1*, pages 3–36. Elsevier, Amsterdam.

Lehmann, E. L. and Romano, J. P. (2022). *Testing Statistical Hypotheses.* Springer, New York, fourth edition.

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature*, 35(1):13–39.

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