Stock Returns and Dividend Yields Revisited: A New Way to Look at an Old Problem

Michael Wolf *
Division of Statistics
UCLA
Los Angeles, CA 90095

July 1997

Abstract

We introduce a new statistical method for finding good confidence intervals for unknown parameters in the context of dependent and possibly heteroskedastic random variables, called subsampling. It works under very weak conditions and avoids the pitfalls of having to choose a structural model to fit to observed data. Appropriate simulation studies suggest that is has better small sample properties than the GMM method, which also works under weak conditions and is model-free. We use the subsampling method to discuss the problem of whether stock returns can be predicted from dividend yields. Looking at three data sets, we do not find convincing evidence for predictability of stock returns.

*I am grateful to Ibbotson and Associates for providing me with data on the S&P 500 index and to Charles Nelson for providing me with data on the NYSE index. I would like to thank Geert Bekaert, William Goetzmann, and Robert Hodrick for helpful comments and suggestions. Also, I got valuable feedback from seminar participants at Stanford, UCLA, UCSB, UCSD, and USC. Any remaining errors are, of course, my own responsibility.

Address for correspondence: Michael Wolf, Phone: (310) 206-3381, Fax: (310) 206-5658, E-mail: mwolf@stat.ucla.edu.

1 Introduction

There has been considerable debate in the recent finance literature whether stock returns can be predicted from dividend yields. Various forms of the so-called efficiency of markets hypothesis imply that all available information of future stock returns is contained in a stock's current price and therefore future returns should be completely unpredictable. However, a number of recent studies appear to provide empirical support for the use of the current dividend-price ratio, the so-called dividend yield, as a measure of expected stock returns. See for example Rozeff (1984), Campbell and Shiller (1988), Fama and French (1988), Hodrick (1992), and Nelson and Kim (1993). The problem with such studies is that stock return regressions face several kinds of statistical problems, among them strong dependency structures and biases in the estimation of regression coefficients. These problems tend to make findings against the no predictability hypothesis appear more significant than they really are.

Having recognized this, Goetzmann and Jorion (1993) argue that previous findings might be spurious and largely due to the bad small sample behavior of commonly used inference methods. They employ a bootstrap approach and conclude that there is no strong evidence indicating that dividend yields can be used to forecast stock returns. One should note, however, that their special approach is not shown to be backed up by theoretical properties. Also, it requires a lot of custom-tailoring to the specific situation at hand. For other scenarios, a different tailoring would be needed.

We intend to help in resolving some of the disagreement by applying a new technique, called subsampling. It has been shown to give correct results under very weak conditions, including dependency and heteroskedasticity. Moreover, it makes use of the observed data in a very intuitive and simple way and does not require any modifications to be applicable in different scenarios. The paper is organized as follows. In Section 2 we give a brief description of the stock returns regression problem as well as a summary of previously used approaches and corresponding findings. Section 3 introduces the proposed subsampling method. Section 4 contains some practical details concerning the actual implementation. We use a simulation study to evaluate small sample properties concerning stock return regressions in Section 5. In Section 6 we apply the subsampling method to three data sets and present the results. Section 7 provides some additional insight dealing with a reorganization of long-horizon returns and a joint test for multiple horizons. The paper ends with some concluding remarks in Section 8.

2 Background and Definitions

We will now describe the stock returns problem in a formal way and look at some of the previous studies in more detail. Most of the empirical studies use monthly data. Define the one-period real total return as $R_{t+1} = (P_{t+1} + d_{t+1})/P_t$, where P_t is the end-of-month real stock price and d_t is the real dividends paid during month t. The total return can be decomposed into capital and income return:

$$R_{t+1} = R^C_{t+1} + R^I_{t+1} \equiv P_{t+1}/P_t + d_{t+1}/P_t. \tag{1}$$

Since dividend payments are highly seasonal, usually a monthly annualized dividend series D_t is computed from compounding twelve monthly dividends at the 1-month Treasury bill rate r_t:

$$D_t = d_t + (1 + r_t)d_{t-1} + (1 + r_t)(1 + r_{t-1})d_{t-2} + \ldots + (1 + r_t)(1 + r_{t-1})\ldots(1 + r_{t-10})d_{t-11}$$

Then annual dividend yield is defined as $Y_t = D_t/P_t$. As a strong form of the efficiency of markets hypothesis, the historic random walk model specifies that the returns R_t are i.i.d. (independent and identically distributed) according to some unknown distribution. Frequently, it is assumed that this distribution is known up to finite number of parameters, that is to say it belongs to a certain parametric family of distributions. Of course, this is done for convenience. In general, one can never verify such a claim. The two most commonly used families for this purpose are the normal and the lognormal families. It seems that the latter is preferred for a number of reason, one of them being that stock returns are basically by definition bounded below by -100%. Indeed, the lognormal random walk model has a long and illustrious history, and has become "the workhorse of the financial asset pricing literature" (Campbell, Lo, and MacKinlay, 1997). One implication of this particular model, but also other forms of the efficiency of markets hypotheses, is that future returns would be completely unpredictable. Especially, a linear regression model like the following

$$\begin{aligned} \ln(R_{t+k,t}) &= \alpha_k + \beta_k(D_t/P_t) + \epsilon_{t+k,k} \\ &= \alpha_k + \beta_k(Y_t) + \epsilon_{t+k,k} \end{aligned} \tag{2}$$

where $\ln(R_{t+k,i}) = \ln(R_{t+1}) + \cdots + \ln(R_{t+k})$ is the continuously compounded k-period return, would have a true β_k coefficient of zero. All of the afore-mentioned studies are concerned with testing the the null hypothesis $H_0 : \beta_k = 0$. Usually a number of return horizons k are considered, since for some theoretical reasons (e.g., present value model) predictability might be suspected to increase with the return horizon. Most studies are able to reject the

null hypothesis at all horizons considered, suggesting that future returns can be partially forecasted using present dividend yields. The empirical evidence is strongest for so-called long horizon returns beyond one year, that is for values $k \geq 12$ when using monthly data. The longest horizon usually considered is four years, or $k = 48$.

It is clear that under the null hypothesis the stochastic behavior of the error variables $\epsilon_{t+k,k}$ in (2) is completely determined by the stochastic behavior of the R_t process. In fact, in that case we have $\epsilon_{t+k,k} = \ln(R_{t+k,i}) - \alpha_k$. Even under the random walk model — which is stronger than the null hypothesis of $\beta_k = 0$ — the $\epsilon_{t+k,k}$ are uncorrelated only for $k = 1$. If the data are sampled more finely than the compound return interval, that is, for $k > 1$, the errors will always exhibit serial correlation due to the resulting overlap. For example, under the random walk model they follow a moving average process of order $k - 1$, or MA$(k - 1)$ process. In case the log returns are correlated, or under the alternative hypothesis, the $\epsilon_{t+k,k}$ can be arbitrarily serially correlated for all values of k. The estimation of β_k can be easily done by ordinary least squares. However, testing the null hypothesis, or assigning a P-value to the observed $\hat{\beta}_k$, is nontrivial for a number of reasons.

1. In the case of correlated residuals, like in the case of long-horizon regressions, the usual OLS standard errors are not valid, since they are based on the hypothesis of uncorrelated residuals.

2. The independent variable in the regression (2) is predetermined but not exogenous. That is to say that Y_t is uncorrelated with the current error term $\epsilon_{t+k,k}$ but generally not with past error terms $\epsilon_{t+k-j,k}$, $j > 1$. This is because

$$\epsilon_{t+k-j,k} = \ln(R_{t+k-j,i}) - \alpha_k - \beta_k(Y_{t-j})$$

   and the dividend yield series Y_t is highly autocorrelated, or highly persistent, at monthly intervals. It is well known that regressions with predetermined independent variables can lead to biased, although consistent, estimates. A standard reference is Stambaugh (1986). In the case of stock return horizons, the OLS estimates of $\hat{\beta}_k$ are typically upward biased.

3. A second source of bias in the OLS estimates is the fact that the regressor behaves like a lagged dependent variable; P_t appears on both the left and right hand side of the regression equation (2).

In the remainder of the paper will discuss various inference methods for β_k according to two criteria: asymptotic consistency and small sample properties.

Loosely speaking, asymptotic consistency means that we get the right answer if the sample size is infinity. For example, if we construct a test for $H_0 : \beta_k = 0$ with nominal significance level $\alpha = 0.05$, then the probability of falsely rejecting H_0 will tend to 0.05, as the sample size tends to infinity. Or, if we construct a confidence interval for the unknown parameter β_k with nominal confidence level $1 - \alpha = 0.95$, then the probability that β_k will be contained in the interval will tend to 0.95, as the sample size tends to infinity. Of course, in practice we never encounter an infinite sample size, but asymptotic consistency is the minimum required from a statistical inference method. If we do not get the right answer in the ideal scenario (infinite sample), then there is no good reason to use in the harsh reality (finite sample). For this reason we should restrict our focus to methods that are asymptotically consistent under sensible conditions.

If for a given problem there was only one asymptotically consistent method, our choice would be easy. However, typically this is not the case and we have to choose between several competing methods. It makes sense then to consider small sample properties, that is, we want to judge how close do various methods get to the right answer when only finite samples are available. The answer to this problem in general must be "it depends". Indeed, small sample properties depend on the true underlying data generating mechanism, which is unknown. Since we therefore can never give a perfect answer, we must look at reasonable and feasible data generating mechanisms that we believe are not too far from the truth. By artificially generating data from known approximating models we can nowadays with the help of fast computers conduct simulation studies, which allow us to gain valuable insight into small sample properties.

Our philosophy is hence the following. For a given problem — in this paper the stock return predictability problem — find inference methods that are asymptotically consistent under not too restrictive conditions. Then distinguish between those methods via appropriate small sample simulation studies, using a reasonable data generating mechanism and real-life sample sizes. Pick the inference method declared the "winner" by the simulation studies as the one which we should entrust our real data the most.

2.1 The GMM Approach

A very common approach for making inference on β_k in the context of dependent and possibly heteroskedastic observations is to correct the standard errors of regression coefficients estimates for serial correlation according to the generalized method of moments (GMM) by Hansen and Hodrick (1980) and Hansen (1982). Most papers that follow this idea base the

correction on the additional hypothesis that log returns are uncorrelated, in which case the residuals of a k-horizon regression follow a simple MA($k-1$) process.

The GMM method fares well in terms of asymptotic consistency. It has been shown to converge to the right answer under weak and very general conditions (see above references).

Small sample properties, on the other hand, might pose a problem. Since GMM uses asymptotic normality, centered at the true β_k, it accounts neither for finite sample biases of $\hat{\beta}_k$ nor for potential skewness to the right of the sampling distribution of $\hat{\beta}_k$ (e.g., Goetzmann and Jorion, 1993). In addition, there is evidence that in the context of serial correlation the GMM corrections of the standard errors are often insufficient in finite samples. For example, see Ferson and Foerster (1994), Bekaert and Urias (1996), and Politis, Romano, and Wolf (1997). We therefore expect the GMM approximation to the true sampling distribution of $\hat{\beta}_k$ to be centered at too small a value and having a right tail that is too short. The consequence is that observed (positive) values of $\hat{\beta}_k$ will be judged as overly significant and hence tests for $\hat{\beta}_k$ will be biased towards false rejection of H_0. These deficiencies of the GMM method were, among others, realized by Goetzmann and Jorion (1993) and Nelson and Kim (1993).

Two examples of studies employing the GMM method can be found in Fama and French (1988) and in Chapter 7 of Campbell, Lo, and MacKinlay (1997). Either study rejects the null hypothesis $\beta_k = 0$ at conventional significance levels, at least for return horizons of one year and beyond.

2.2 The VAR Approach

An alternative approach is to estimate the sampling distribution of $\hat{\beta}_k$ *under the null hypothesis*, and to use this estimated distribution to attach a P-value to the observed value of $\hat{\beta}_k$. The typical way of estimating the null sampling distribution involves simulating artificial return and dividend yield sequences, employing a data generating mechanism which imposes the null hypothesis. A large number, B say, of such simulations are carried out. For each outcome, the corresponding estimate $\hat{\beta}_k^*$ is computed. The empirical distribution of the B $\hat{\beta}_k^*$ values then serves as the desired estimate of the sampling distribution of $\hat{\beta}_k$ under the null. A one-sided P-value is given by the proportion of $\hat{\beta}_k^*$ values that exceed the observed statistic $\hat{\beta}_k$. For a general reference on this idea see Noreen (1989). There have been two different appropriate data generating mechanisms for stock return sequences proposed in the literature.

Campbell and Shiller (1989), Hodrick (1992), Nelson and Kim (1993), and Goetzmann and Jorion (1995), among others, consider a first-order vector autoregression (VAR) in at least the following two variables: log return and dividend yield. Sometimes, additional variables are included. For example, Campbell and Shiller (1989) include a term corresponding to earnings price ratio. Hodrick (1992) includes the one-month Treasury-bill return relative to its previous 12-month moving average which is denoted rb_t. To describe his model, say, let

$$Z_t \equiv [\ln(R_t) - E(\ln(R_t), D_t/P_t - E(D_t/P_t), rb_t - E(rb_t)]^T.$$

Then a first order VAR, or VAR(1), is given by

$$Z_{t+1} = AZ_t + u_{t+1}, \tag{3}$$

where A is a 3 x 3 matrix and u_t is a 3-dimensional white noise innovation sequence.

Hodrick (1992) fits this model to the observed data and then sets the row of the estimated VAR(1) matrix corresponding to log returns to zero, and the constant term corresponding to log returns to the unconditional mean implied by the original VAR. Of course, specifying the VAR parameters is not sufficient, as an innovation sequence u_t has to be fed to the VAR model. Since there is strong empirical evidence for return data to exhibit (conditional) heteroskedasticity, Hodrick fits a generalized autoregressive conditionally heteroskedastic (GARCH) model to the innovations estimated by the VAR. For more details the reader is referred to the original paper. He then generates artificial innovation sequences according to the estimated GARCH process, where the innovations have a conditional normal distribution. Using this approach, Hodrick also finds evidence of predictability in stock returns, both for short and long horizons.

Nelson and Kim (1993) employ a similar method, simulating from a VAR model under the null hypothesis. However, they randomize fitted innovations for the artificial innovation sequences in order to better match the dispersion of true marginal distribution of the innovations. The disadvantage of this method is that it destroys any potential dependence in the innovation sequence. The study reports that the simulated distributions of the regular t-statistics are upward biased and that these biases should be taken into account when making inference. However, even after a bias correction, the authors find some evidence for predictability, especially when looking at post-war data after 1947.

Unlike the GMM method, the VAR approach tries to capture the finite sample distribution of $\hat{\beta}_k$ by generating artificial data having the sample size as the observed data. It succeeds in correcting for both upward biases and skewness to some extent, as demonstrated in Nelson and Kim (1993) and Goetzmann and Jorion (1993). However, for many

financial data, using GARCH innovations with a conditional normal distribution (Hodrick, 1992) tends to underestimate the tails of the true sampling distribution. See Remark 5.1 for evidence on this claim. Underestimating the tails will result in overstating the significance of observed $\hat{\beta}_k$ values again. This might explain why the finding of Nelson and Kim (1993) who randomize fitted innovations are not as significant as those of Hodrick (1992). On the other hand, the small sample effect of destroying the correlation in the second moments of the innovations is not clear.

The obvious shortcoming of the VAR approach is the use of a structural model. Asymptotic consistency will only be assured if VAR(1) is the true model. This is doubtful. Of course, how big the asymptotic mistake is depends on how far the true mechanism is away from VAR(1). The problem is magnified if a parametric model for the innovations, such as GARCH(1,1), is used. In addition, it is noteworthy that the VAR model is estimated from monthly, nonoverlapping data. Small mistakes for $k = 1$ will therefore be magnified for long horizons, such as $k = 48$, via adding up k one-month returns to construct a k-month return.

Note that is very awkward to judge the small sample properties of the VAR method via simulation studies. Hodrick (1992) presents a simulation study that paints a very favorable picture. The problem is that he uses VAR(1) with GARCH(1,1) innovations as the data generating mechanism in the study, that is, he pretends to know what the true mechanism is. Such a study is bound to be overly optimistic. To make a simple analogy, one should not judge the small sample properties of the t-test by generating i.i.d. normal observations only. On the other hand, one could easily arrive at an overly pessimistic answer by employing a data generating mechanism that is far away from VAR(1) in some sense.

2.3 A Bootstrap Approach

As an alternative to the VAR method, Goetzmann and Jorion (1993) use a bootstrap approach to generate artificial data sequences under the null hypothesis. The motivation is that a model-free method such as the bootstrap will avoid any mistakes due a potentially misspecified structural model. Their particular bootstrap works as follows.

1. Form the empirical distribution of monthly total stock returns R_t and their associated income returns R_t^I, as defined in (1), from the observed data.

2. Generate a pseudo return sequence R_t^* i.i.d. according to the empirical distribution of the observed total returns $R_1 \ldots R_n$.

3. Subtract the contemporaneous income returns $R_t^{I,*}$ to create a pseudo capital-return series $R_t^{C,*}$: $R_t^{C,*} = R_t^* - R_t^{I,*}$. Compound these to create a pseudo price series P_t^*.

4. Create a pseudo dividend yield sequence $Y_t^* = D_t/P_t^*$, in which the D_t are the actual annual dividend flows.

It is obvious that some custom-tailoring is employed here in the attempt of capturing the relationship between price levels and dividends. The key problem with this approach is seen in the fact that total returns are resampled at random, implying that returns are i.i.d. according to some unknown distribution, while dividend flows remain fixed, implying that dividend payments are completely nonstochastic. While the first assumption is slightly troublesome — the null hypothesis of a random walk is stronger than the null hypothesis of no predictability — the second assumption seems unrealistic. Clearly, we do not for certain what the dividend payments of a certain stock, or stock index, will be in the future. At least to some extent, these payments will be determined by random factors, such as interest rates, state of the overall economy, and events associated with the particular companies. For this reason, the asymptotic consistency of this bootstrap approach is doubtful. Goetzmann and Jorion (1993) do not discuss the asymptotic properties of their method.

Note that Goetzmann and Jorion come to basically opposite conclusions as all previous studies. They do not find strong statistical evidence in favor of predictability of stock returns. P-values of the observed $\hat{\beta}_k$ values are typically slightly above 10%, even for long-term horizons.

3 A New Approach

It has been almost two decades since Efron (1979) introduced the bootstrap procedure for estimating sampling distributions of statistics based on independent and identically distributed (i.i.d.) observations. In practice, the bootstrap method generates so-called pseudo sequences, or artificial sequences, by selecting single data points at random from the observed sequence and joining them together. Then the statistic of interest (mean, variance, regression coefficient, etc.) is computed from the pseudo sequence, yielding a pseudo statistic. This process is repeated a large number of times and the empirical distribution of all individual pseudo statistics is used to find an approximation to the unknown sampling distribution of the original statistic. A very nice introduction to the bootstrap can be found in Efron and Tibshirani (1993).

It is well known that the bootstrap often gives more accurate approximations than classical large sample approximations, like normal approximations. However, when the observations are dependent this "classical" bootstrap no longer succeeds. Singh (1981) showed that Efron's bootstrap fails to capture the dependence structure even for the sample mean of m-dependent data. Following this observation, there have been several attempts to modify and extend Efron's idea to dependent data.

There are, broadly speaking, two approaches to using resampling methods for stationary dependent data. One is to apply Efron's bootstrap to an approximate i.i.d. setting by focusing on the residuals of some general regression model. Such examples include linear regression (Freedman, 1981; Freedman, 1984; Wu, 1986; Liu, 1988), autoregressive time series (Efron and Tibshirani, 1986; Bose, 1988), nonparametric regression and nonparametric kernel spectral estimation (Härdle and Bowman, 1988; Franke and Härdle, 1992). In all of the above situations the residuals are resampled, not the original observations. However, this method is restricted to relatively simple contexts where structural models are both plausible and tractable. As a second approach, resampling methods for less restrictive contexts have been suggested more recently. They are based on so-called "blocking" arguments, in which the data are divided into blocks and these blocks, rather than individual data values or estimated residuals, are resampled and joined together to form new pseudo time series. Carlstein (1986) proposed non-overlapping blocks, whereas Künsch (1989) and Liu and Singh (1992) independently introduced the 'moving blocks' method which employs overlapping blocks. Subsequent research seems to have favored this scheme.

As an alternative to bootstrap methods, Politis and Romano (1994) proposed the subsampling approach. Rather than resampling blocks from the original time series as ingredients to generating new pseudo time series, each individual block of observations is looked upon as a valid 'subseries' in its own right. The motivation is as follows. Each block, as a part of the original series, was generated by the true underlying probability mechanism. It then seems reasonable to hope that one can gain information about the sampling distribution of a statistic — like the least squares estimator $\hat{\beta}_k$ — by evaluating it on all subseries, or 'subsamples'. An attractive feature of the subsampling method is that it has been shown to be asymptotically consistent under very weak assumptions. For example, it can handle many of the counterexamples of the bootstrap. Apart from regularity and dependency conditions, the only requirement, in the stationary setup, is that the sampling distribution of the properly normalized statistic of interest has a nondegenerate limiting distribution. The moving blocks method, on the other hand, has essentially been shown to be valid for functions of linear statistics and smooth functionals only (see Künsch (1989)

and Bühlmann (1994)). Note that OLS estimators are linear statistics and hence could be handled by the moving blocks bootstrap as well.

An extension of the subsampling method to the heteroskedastic, or nonstationary, case was presented in Politis, Romano, and Wolf (1997). This is an important result, since many interesting time series are known to exhibit some kind of heteroskedasticity.

We will proceed to give a brief description of the subsampling method here in order to make this paper self-contained. For more details and a derivation of theoretical properties, the reader is referred to Politis and Romano (1994) and Politis, Romano, and Wolf (1997).

Suppose $\{\ldots, X_{-1}, X_0, X_1, \ldots\}$ is a sequence of real or vector-valued random variables, defined on a common probability space and governed by a joint probability law P. The goal is to construct a confidence interval for some real-valued parameter $\theta = \theta(P)$, on the basis of observing a finite sample $\{X_1, \ldots, X_n\}$. For example, in the case of the return regression (2), X_t is equal to the vector $(\ln(R_{t+k,t}), (D_t/P_t))$ and θ is equal to the regression coefficient β_k. Note that this theory can be generalized to multivariate or even functional parameters. The time series X_t is assumed to satisfy a certain weak dependence condition, the so-called strong mixing condition. This is a sufficient condition to ensure that we gain enough additional information from a bigger sample and hence asymptotic theory go through. Processes used to model financial time series, such as autoregressive processes, typically satisfy the strong mixing condition.

Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be an estimator of θ and $\hat{\theta}_{b,a} = \hat{\theta}_b(X_a, \ldots, X_{a+b-1})$ be the estimator of θ based on the smaller block, or subsample, $X_a, \ldots, X_{a+b-1}$. The block size b will be much smaller than the sample size n. Define $J_{b,a}$ to be the sampling distribution of $\tau_b\ (\hat{\theta}_{b,a} - \theta)$, where τ_b is an appropriate normalizing constant. The corresponding cumulative distribution function is then

$$J_{b,a}(x) = Prob\{\tau_b(\hat{\theta}_{b,a} - \theta) \leq x\}. \tag{4}$$

The normalizing constant τ_b is needed to put the differences $\hat{\theta}_{b,a} - \theta$ "on the same scale" as the difference $\hat{\theta}_n - \theta$, since block estimates are based on a smaller sample size than the estimate from the entire sample. Loosely speaking, the normalizing constant ensures that both $J_{b,a}$ and $J_{n,1}$ have a nondegenerate limiting distribution. In most cases the proper normalizing constant is simply the square root of the corresponding sample size, that is, $\tau_b = b^{1/2}$ and $\tau_n = n^{1/2}$. But, other cases also exist; e.g., see Politis and Romano (1994),

We want to estimate the sampling distribution of our estimator $\hat{\theta}_n$, properly standardized, namely

$$J_{n,1}(x) = Prob\{\tau_n(\hat{\theta}_n - \theta) \leq x\}. \tag{5}$$

Indeed, if we knew it exactly, we could construct exact confidence intervals for θ. Since this is the case in only the rarest of instances, we usually have to rely on some sort of approximation. Typically, the approximations will become better and better as the sample size increases. The most common approximation is the normal approximation, utilizing the fact that in many scenarios $J_{n,1}$ has a limiting normal distribution. Aided by modern computer technology, statisticians have during the last twenty years developed computationally intensive methods which for small samples often provide better approximations than the normal one.

In order to describe the subsampling method, let $Y_{b,a}$ be the block of size b of the consecutive data $\{X_a, \ldots, X_{a+b-1}\}$. The observed sequence $X_1, \ldots, X_n$ yields $n-b+1$ such blocks. The first one is $Y_{b,1} = \{X_1, \ldots, X_b\}$, the last one is $Y_{b,n-b+1} = \{X_{n-b+1}, \ldots, X_n\}$. As defined before, $\hat{\theta}_{b,a}$ is equal to the statistic $\hat{\theta}_b$ evaluated at the data set $Y_{b,a}$. The subsampling approximation to $J_{n,1}(x)$ is now given by

$$L_n(x) = \frac{1}{n-b+1} \sum_{a=1}^{n-b+1} 1\{\tau_b(\hat{\theta}_{b,a} - \hat{\theta}_n) \leq x\}, \tag{6}$$

where $1\{\cdot\}$ is the indicator function. Therefore, the subsampling approximation is simply the proportion of properly standardized subsample statistics less than or equal to x.

The motivation behind the method is the following. For any a, $Y_{b,a}$ is a 'true' subsample of size b. Hence, the *exact* distribution of $\tau_b(\hat{\theta}_{b,a} - \theta)$ is $J_{b,a}$. For large sample sizes, $J_{b,a}$ will be close to $J_{n,1}$ for all indices a, at least in the stationary setup. In the context of heteroskedastic observations the same will hold true under some regularity conditions (see Politis, Romano, and Wolf, 1997). The empirical distribution of the $n-b+1$ values of $\tau_b(\hat{\theta}_{b,a} - \theta)$ should then serve as good approximation to $J_{n,1}$. The fact that we replace θ by $\hat{\theta}_{n,1}$ in (6) does not affect the asymptotic properties, given some very weak conditions on the block size b. Using the approximation (6) allows us to construct one-sided confidence confidence intervals for θ in the following way. A one-sided lower $1-\alpha$ interval is given by

$$I_{LOW} = [\hat{\theta}_{n,1} - \tau_n^{-1} c_n(1-\alpha),\ \infty) \tag{7}$$

and a one-sided upper $1-\alpha$ interval is given by

$$I_{UP} = (-\infty,\ \hat{\theta}_{n,1} - \tau_n^{-1} c_n(\alpha)], \tag{8}$$

where $c_n(\lambda)$ denotes a λ quantile of the subsampling distribution L_n defined in (6). Two-sided confidence intervals can be constructed as the intersection of two one-sided intervals. For example, the intersection of a lower and an upper one-sided 95% interval yields a two-sided 90% confidence interval. Such intervals are called *equal-tailed* because they have approximately equal probability in each tail. As an alternative approach two-sided *symmetric* confidence intervals can be constructed. Their name stems from the fact that they extend equally far to the left as to the right of the estimate $\hat{\theta}_n$, just like normal intervals do. The common way to construct symmetric confidence intervals is to estimate the two-sided cumulative distribution function

$$J_{n,1,|\cdot|}(x) = Prob\{\tau_n \left|\hat{\theta}_{n,1} - \theta\right| \leq x\}. \tag{9}$$

The subsampling approximation to $J_{n,1,|\cdot|}(x)$ is defined by

$$L_{n,|\cdot|}(x) = \frac{1}{n-b+1} \sum_{a=1}^{n-b+1} 1\{\tau_b \left|\hat{\theta}_{b,a} - \hat{\theta}_{n,1}\right| \leq x\}. \tag{10}$$

A two-sided symmetric $(1-\alpha)$ confidence interval is then given by

$$I_{SYM} = [\hat{\theta}_{n,1} - \tau_n^{-1} c_{n,|\cdot|}(1-\ \alpha),\ \hat{\theta}_{n,1} + \tau_n^{-1} c_{n,|\cdot|}(1-\ \alpha)], \tag{11}$$

where $c_{n,|\cdot|}(\lambda)$ denotes a λ quantile of the subsampling distribution $L_{n,|\cdot|}$ defined in (10). Why is it useful to distinguish between equal-tailed and symmetric intervals? It is known that symmetric intervals often enjoy enhanced coverage properties and, even in asymmetric circumstances, can be shorter than equal-tailed intervals (e.g., Hall, 1988). Some corresponding simulation studies can be found in Politis, Romano, and Wolf (1997) and Wolf (1996). We will use symmetric subsampling intervals for the remainder of this paper.

It can be shown that under very weak conditions the subsampling method will yield asymptotically correct inference. As far as confidence intervals are concerned, this means that as the simple size n increases to infinity the actual coverage probability will tend to the nominal level $1-\alpha$. The only conditions needed to assure this property are the existence of a nondegenerate limiting distribution for $J_{n,1}$, a universal moment bound on the X_t, a certain mixing condition, a bound on the amount of global heteroskedasticity, and some requirements on the block size b. Detailed theorems can be found in Politis, Romano, and Wolf (1997).

Remark 3.1 An important advantage of the subsampling method is the fact that is enough to know about the existence of a limiting distribution. It does not have to be estimated in

practice. Numerous examples exist where the limiting distribution depends in a complicated way on the underlying data generating mechanism, making inference very difficult or even impossible if explicit estimation of this distribution is necessary. One example is the area of variance ratio tests, where the estimation of the limiting variance of the test statistic is usually done under simplifying assumptions (e.g., Lo and MacKinlay, 1988). Another example is given in Subsection 7.2, where the goal is to make joint inference for a number of return horizons. Some non-finance examples are discussed in Politis and Romano (1994).

A practical problem lies in choosing the block size b. To ensure the asymptotic properties of the method it is only necessary that the block size b tend to infinity with the sample size n, but at a smaller rate: $b \to \infty$ and $b/n \to 0$, as $n \to \infty$. Of course, this rule gives us very little guidance for applications. Usually, we have a sample of fixed size n. Picking $b = n^{1/2}$ would be consistent with the above condition, but many other choices would be as well. As to be expected, small sample properties depend on the actual choice of b (Politis, Romano, and Wolf, 1997; among others), with the dependency diminishing as the sample size increases.

In some sense, the block size b might be called a hidden parameter, or model parameter. Having to choose such a hidden parameter is a property that the subsampling method shares with many other statistical inference methods. For example, for density estimation, typically a bandwidth parameter has to be selected. When using GMM in the context of dependent observations, again a bandwidth parameter has to be selected in using a kernel for estimating the limiting covariance matrix. This problem sometimes seems to be swept under the rug in the applied literature. Fama and French (1988) and Campbell et al. (1997) use the Hansen and Hodrick (1980) kernel, weighting autocovariances up to lag $k-1$ with weight one and autocovariances beyond lag k with weight zero. This will produce a consistent estimator of the limiting covariance matrix only under the null hypothesis of $\beta_k = 0$ *and* under the additional assumptions that the log returns are uncorrelated. If one is interested in finding confidence intervals for β_k or suspects returns to be correlated (there is evidence for small, but significant, correlation at the monthly horizon), a more general kernel has to be used. In that case, the choice of the bandwidth is not obvious. Andrews (1991) compares a number of kernels and suggests an automatic bandwidth selection procedure based on asymptotic considerations. He finds that the so-called Quadratic Spectral (QS) kernels dominates other kernels, both in terms of asymptotic theory and small sample simulation studies. We will use this kernel, in conjunction with Andrew's bandwidth selection procedure, when comparing GMM to subsampling in our simulation studies.

To deal with the problem of choosing the block size b for the subsampling method we suggest a calibration technique which in a sense avoids having the find the "best" block size.

4 Calibration

One can think of the accuracy of an approximate or asymptotic confidence procedure — such as normal, bootstrap, or subsampling methods — in terms of its *calibration* (Loh, 1987). Suppose we use the procedure to construct a confidence interval with nominal confidence level $1-\lambda$. We can denote the actual confidence level by $1-\alpha$. λ is known to us, α typically is not. An asymptotic method only ensures that $1-\lambda$ will tend to $1-\alpha$ as the sample size tends to infinity. For a finite sample size, the two levels might not be the same. If we knew the calibration function $h : 1-\lambda \to 1-\alpha$, we could construct a confidence region with exactly the desired coverage, by selecting the value of λ that satisfies $h(1-\lambda) = 1-\alpha$. For example, if $h(0.98) = 0.95$, then a confidence interval with nominal level 98% would be an actual 95% confidence interval.

Fortunately, the calibration function $h(\cdot)$ can be estimated. One way of doing this would be to assume a parametric model with a known parameter θ_0. By then using a Monte Carlo approach a natural estimate $\hat{h}(\cdot)$ would be easy to find: One would generate many artificial sequences, compute a $1-\lambda$ interval for each sequence, and take the proportion of intervals that contain θ_0. Of course, if we were willing to assume a parametric model, why use a model-free technique such as the subsampling method in the first place? It is therefore more desirable to use a model-free data generating mechanism for the Monte Carlo approach. The obvious choice is to use the bootstrap. We therefore generate artificial sequences from a bootstrap distribution P_n^*, then construct a confidence interval from each generated pseudo sequence, and observe how frequently the parameter $\hat{\theta}_n$ is contained in those intervals. In the context of dependent data, we need to employ a bootstrap suitable for time series. The moving blocks bootstrap (Künsch, 1989), which was briefly mentioned at the beginning of Section 3, lends itself to the task. It generates pseudo sequences $X_1^*, \ldots, X_n^*$ by resampling entire blocks from the original data and joining these together, rather than single data points. Formally, let $Y_{b,a}$ be the block of size b of the data $\{X_a, \ldots, X_{a+b-1}\}$. For simplicity we assume that $n = lb$, for some integer l. Also, let P_n^* denote the empirical distribution of the blocks $Y_{b,1}, Y_{b,2}, \ldots, Y_{b,n-b+1}$. Then a pseudo sequence is constructed by choosing $Y_{b,1}^*, \ldots, Y_{b,l}^*$ i.i.d. from P_n^* and concatenating them. In case n is not a multiple of b, we

use the same algorithm with the smallest l for which $n < lb$ and truncate the so-obtained sequences at n.

In case we want to apply the calibration scheme to the subsampling method, we can do it conditional on a *reasonable* block size. This means that we fix a sensible block size and calibrate the subsampling intervals using that particular block size. This eliminates the problem of finding the "best" block size. In some scenarios we will have a pretty good idea what a reasonable block size will be, either from prior experience or related simulation studies. Otherwise, see Remark 4.1 below. To describe the calibration technique more formally we can use the following algorithm.

Description of the Calibration Method:

1. Generate K pseudo sequences $X_1^{*k}, \ldots, X_n^{*k}$, according to a moving blocks bootstrap distribution P_n^*.
   For each sequence, $k = 1, \ldots, K$,

   1a. Compute an $1-\lambda$ level confidence interval $\mathrm{CI}_{1-\lambda}^k$, for a grid of values of λ in the neighborhood of α.

2. For each λ compute $\hat{h}(1-\lambda) = \#\{\hat{\theta}_n \in \mathrm{CI}_{1-\lambda}^k\}/K$.

3. Interpolate $\hat{h}(\cdot)$ between the grid values.

4. Find the value of λ satisfying $\hat{h}(1-\lambda) = 1-\alpha$.

5. Construct a confidence interval with nominal level $1-\lambda$.

Remark 4.1

1. The moving blocks bootstrap in step 1 of the above algorithm requires its own block size b_{MB}. The choice of this block size has a second order effect and is therefore not very important. However, if an automatic selection method is preferred, a "nested bootstrap" can be used. That means that we would use the moving blocks bootstrap in both steps 1 and 1a of the above algorithm with the same block size $b_M B$, limiting the grid of λ values to $\lambda = \alpha$. Repeating this algorithm for a number of b_{MB} values, we then would select the value b_{MB} which yields estimated coverage closest to $1-\alpha$.

2. If we use the calibration scheme to calibrate the subsampling method we need to start out with a reasonable block size b. In situations where we do not know what a reasonable block size is, we can use the following idea. In the same way as the actual confidence level can be regarded as function of the nominal confidence level (conditional on a fixed block size), it can be considered as a function of the block size (conditional on a fixed nominal level). Fixing the nominal level at the desired level, that is, choosing $\lambda = \alpha$, we can therefore estimate the block calibration function $g : b \to 1 - \alpha$, using an analogous calibration algorithm:

   1* Generate K pseudo sequences $X_1^{*k}, \ldots, X_n^{*k}$, according to a moving blocks bootstrap distribution P_n^*.
   For each sequence, $k = 1, \ldots, K$,

   1a*. Compute an $1 - \alpha$ level confidence interval CI_b^k, for a selection of block sizes b.

   2* For each b compute $\hat{g}(b) = \#\{\hat{\theta}_n \in \mathrm{CI}_b^k\}/K$.

   A reasonable block size will then satisfy $\hat{g}(b) \simeq 1 - \alpha$.

3. Two-sided equal-tailed intervals should always be computed as the intersection of two separately calibrated one-sided intervals. Particularly if the sampling distribution of $\hat{\theta}_n$ is asymmetric, the amount of calibration needed in the lower tail can be very different from the one needed in the upper tail.

As an illustration of how we would use the calibration method see Figure 1 for a real-life example. The goal is to construct a 95% confidence interval for β_{12} for the S&P 500 post-war data, starting December 1947. We estimate the calibration function $h(\cdot)$ at the discrete points $0.90, 0.91, \ldots, 0.99$, and linearly interpolate in between. Our estimate tells us that we should construct a confidence interval using a nominal level of 0.96. In the calibration algorithm we used $b_{MB} = 100$ in step 1. and $b = 60$ in step 1a.. The latter was estimated according to Remark 4.1, item 2..

Since the calibration is based on the estimated calibration function $\hat{h}(\cdot)$ rather than the true function $h(\cdot)$, calibrated intervals still are not exact. It can be easily shown that calibrating an asymptotically correct procedure, such as subsampling confidence intervals, results in an asymptotically correct procedure again. In the context of i.i.d. random variables it is known that calibrated confidence intervals have better asymptotic properties than uncalibrated intervals (e.g., Efron and Tibshirani, 1993). In technical terms, calibrated confidence intervals for i.i.d. data are generally second-order correct versus first-order correct.

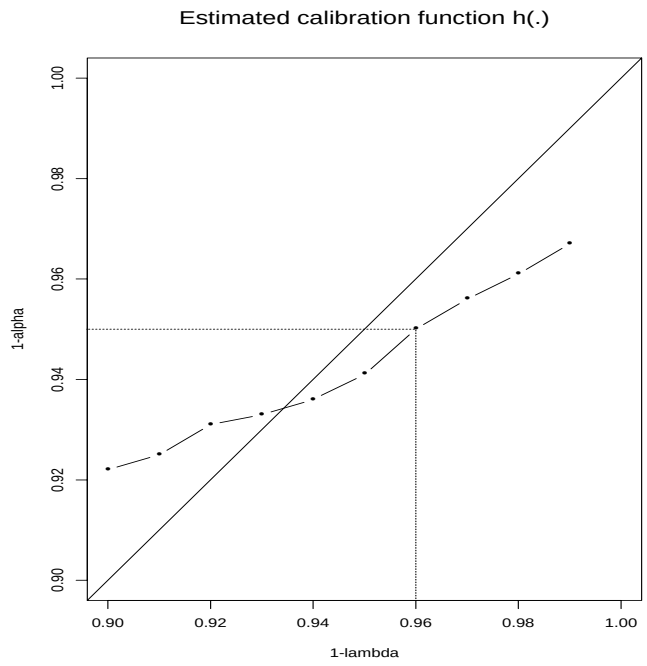

Figure 1: S&P 500 post-war data, return horizon is $k = 12$. Estimates of $h(1-\lambda)$ for a grid of λ values $\{0.01, 0.02, \ldots, 0.1\}$ with linear interpolation in between. The estimate of the calibration function $h(\cdot)$ tells us that if we want a confidence interval for β_{12} with actual level $1-\alpha = 0.95$ we should use a nominal level $1-\lambda = 0.96$.

Corresponding results for dependent data are work in progress and have not yet been obtained. Simulation studies support the conjecture that also in the context of dependent data calibrated intervals generally give better results; see Politis, Romano, and Wolf (1997).

To be fair, calibration techniques can be used to potentially enhance any asymptotic method. An obvious idea in the context of stock return regressions would be to to use a similar calibration method as the above algorithm for GMM confidence intervals. It is the purpose of this paper, however, to compare our new results with previous results in the literature. Therefore, we use the use the "simple" GMM method, as employed by Fama and French (1988) and Campbell et al. (1997), in our simulation studies.

5 A Small Sample Comparison

As noted before, we use two criteria to judge inference methods for β_k, asymptotic consistency and small sample properties. We already mentioned that both GMM and subsampling

give asymptotically correct results under reasonable conditions, while VAR and the Goetzmann and Jorion (1993) bootstrap only work under restrictive conditions. In this section we compare the small sample properties of the GMM and subsampling via simulation studies. We do not include VAR in the studies, since it is based on a structural model. Through a simulation study we could therefore make it look arbitrarily good — by choosing the data generating mechanism equal to the VAR model — or arbitrarily bad — by choosing a data generating mechanism very different from the VAR model. The Goetzmann and Jorion (1993) bootstrap is not included, since it needs actual dividend flows for which we could not think of a good data generating mechanism. Also, since this bootstrap makes the doubtful assumption that dividend flows are nonstochastic, it is not really a competitor for GMM and subsampling on asymptotic grounds.

For our simulations we need a data generating mechanism which jointly models log returns and dividend yields. While we will never know the true mechanism that yielded the observed data, we aim for a reasonable approximation that includes at least two important features. On the one hand, the bias of $\hat{\beta}_k$ due to the predetermined predictor, and on the other hand, the increasing autocorrelation of the residuals with the return horizon k. Both features are captured by the VAR model. Note that it is not a contradiction to employ a model for a *simulation study* which we earlier criticized when used for making *inference*. GMM as well as subsampling are model-free inference methods and do not exploit the dubious knowledge of a structural model. A misspecification of the model has much less impact in a simulation study than when used for making inference. To give an example, we use VAR with GARCH innovations for our simulation to capture the correlation of the second moments of the fitted innovations. This leaves us to make a choice for the conditional distribution of the GARCH model. Two choices are normal and t with a small number of degrees of freedom. In the latter case, the tails are heavier. This has a big effect when making inference, that is, judging the significance of observed statistics. For a simulation study, however, the impact is negligible.

To be specific we will use a first-order VAR model as our data generating mechanism, jointly modeling log return and dividend yield as the vector $X_t \equiv (\ln(R_t), D_t/P_t)$. Let

$$Z_t \equiv [\ln(R_t) - E(\ln(R_t)), D_t/P_t - E(D_t/P_t))]^T.$$

Then our VAR(1) is given by

$$Z_{t+1} = AZ_t + u_{t+1}, \tag{12}$$

where A is a 2 x 2 matrix and u_t is a white noise innovation process. We fit this model to the observed data by least squares and then impose the null hypothesis by setting the first

row of A equal to zero. Since we are concerned with a simulation study only, we do not have to worry about the overall mean and can set it equal to zero without loss of generality. As a simulation study involving calibrated intervals is computationally very expensive, we only consider the shorter post-war data sets, where the case of predictability seems somewhat stronger anyway. We look at three different data sets, the NYSE equal-weighted and value-weighted indices and the S&P 500 index, all starting in December 1947. Both of the NYSE data sets consist of 480 basic observations (12/1947 to 12/1986), the S&P 500 data set consists of 577 observations (12/1947 to 01/1995). The fitted VAR parameters for the data sets under consideration are presented in Table 1.

To generate artificial X_t^* sequences we need vector innovation sequences u_t^*. We use the the constant correlation model, which was introduced by Bollerslev (1990). Namely, let $H_t = E_t(u_{t+1}u_{t+1}^T)$ be the conditional covariance matrix of the first order VAR in (12) with typical element $h_{ij,t}$. Both conditional variances follow an ARMA(1,1) process:

$$h_{ii,t} = \omega_i + \beta_i h_{ii,t-1} + \alpha_i u_{i,t}^2, \quad i = 1, 2. \tag{13}$$

To model the covariance of H_t, the six parameters of (13) are estimated simultaneously with a constant correlation coefficient ρ_{12} by maximum likelihood, assuming conditional normality; see Bollerslev (1990) for more details. The parameter estimates for our three different data sets are reported in Table 2. To judge the size of the ω parameters we should mention that we fitted the models on the percentage scale, that is, a typical monthly return was on the order of 0.5 to 0.8 rather than 0.005 to 0.008. Of course, the α and β parameters do not depend on the choice of scale.

Artificial innovation sequences u_t^* as input to the VAR model (3) are then generated by computing the Cholesky decomposition of the conditional covariance matrix, $C_t^T C_t = H_t$, and setting $u_{t+1}^* = C_t^T \epsilon_{t+1}$, where ϵ_t is a sequence of independent bivariate standard normal random variables. Note that when actually generating those sequences we discard the first 100 observations to avoid start-up effects.

Long horizon return data X_t^* can now be created by feeding the artificial innovation sequences u_t^* into the fitted VAR models, after imposing the null hypothesis by setting the first row of the VAR matrices equal to zero. We also discard the first 100 observations in this step. Finally, the artificial long horizons returns are compounded according to the formula $\ln(R_{t+k,i}^*) = \ln(R_{t+1}^*) + \cdots + \ln(R_{t+k}^*)$. As said before, we only consider the post-war case for the simulation study. When generating the artificial data we obviously need to match the original sample sizes, which are bigger for the S&P 500 data. For every scenario we generate 1000 artificial sequences and compute a 95% calibrated subsampling

interval for β_k for each sequence. For comparison, we also compute confidence intervals using the GMM method, employing the so-called Quadratic Spectral kernel. This kernel was found to have some optimality properties by Andrews (1991). The bandwidth for the kernel was chosen according to the automatic selection procedure of Andrews (1991). We report the percentage of intervals which contain the true parameter zero in Table 3. Note that we carried out the same simulations using conditional innovations having a (scaled) t-distribution with 4 degrees of freedoms. The results were essentially the same and are therefore not reported.

One can see that, except for $k = 1$, the GMM intervals undercover consistently. In other words, the GMM method is biased towards falsely rejecting the null hypothesis. For long horizons of $k = 36$ and $k = 48$, the estimated coverage can be off by almost 20%. On the other hand, the subsampling intervals perform very well for $k \leq 24$. If anything, they tend to overcover somewhat. However, the strong dependency structures of long horizon regressions also cause the subsampling intervals to undercover, though far less so than the GMM intervals. For $k = 36$, the estimated coverage is off between 2% and 3%; for $k = 48$, it is between 4% and 7%. Based on these simulation studies, the small sample properties of the subsampling method appear superior to those of the GMM. This is consistent with previous simulation studies concerning regression coefficients in the context of dependent observations. See Politis, Romano, and Wolf (1997) for a related study employing some different data generating mechanisms and obtaining similar results.

Remark 5.1 In Section 2 we commented on the danger of simulating from a VAR model using GARCH innovation sequences in order to compute a P-value for an *observed* statistic such as $\hat{\beta}_k$. Even in case the fitted VAR model is a good approximation, if the tails of the artificial GARCH sequences are too light, then one overestimates the significance of observed statistics. As a quick check for such a violation, it is interesting to compare large quantiles of the fitted innovations of the three VAR models from Table 1 with the matching quantiles from the corresponding GARCH models from Table 2. We approximated the sampling distribution of 95% and 99% quantiles estimated from GARCH innovation sequences by generating 1000 GARCH sequences of the proper length for each of the three models. There are six marginal distributions altogether, since each of the three VAR models has a two-dimensional innovation sequence. The first dimension corresponds to log return, while the second dimension corresponds to dividend yield. Table 4 presents the results concerning the 95% quantile. It compares the observed statistic computed from the fitted innovations with the sampling distribution of the statistic when the innovations follow the estimated

GARCH process. The sampling distribution is characterized by the mean, the median, and the 1% and 99% quantiles (based on 1000 repetitions). The results concerning the 99% quantile of the innovations were very similar, with identical P-values, and are therefore not presented.

Except for the log return innovations of the model for the S&P 500 data, all 95% quantiles of the fitted innovations are much too big to be compatible with the corresponding GARCH distribution. The two-sided P-values, based on the percentage of simulated 95% quantiles less than or greater than the observed statistic, are equal to zero. For the log return innovations of the model for the S&P 500 data, the observed quantile is too small; again the two-sided P-value is equal to zero. Clearly, this is evidence that using conditionally normal GARCH innovation sequences tends to underestimate the tails of the true sampling distribution and can give misleading results when used to assess the significance of observed statistics. In addition, it appears that for multivariate models one should model each dimension separately.

6 A New Look at Return Regressions

Given the previous discussions, it seems worthwhile to apply the subsampling methodology to stock return regression. Unlike the VAR method and the Goetzmann and Jorion (1993) bootstrap approach it is asymptotically consistent under reasonable conditions. On the other hand, our simulation study indicates that it has better small sample properties than the GMM method employed by Fama and French (1988) and Campbell et al. (1997).

We use three different data sets which have been previously analyzed in the literature. Fama and French (1988) and Nelson and Kim (1993) report regressions of log returns for value-weighted and equally weighted stock portfolios based on the CRSP files for NYSE stocks. Goetzmann and Jorion (1993) use monthly data on the S&P 500 index. In accordance with the majority of the literature, we consider return horizons of 1, 12, 24, 36, and 48 months. Both of the NYSE data sets consist of 721 basic observations (12/1926 to 12/1986), the S&P 500 data set consists of 818 observations (12/1926 to 01/1995). Notice that for return regressions the sample size is reduced by the return horizon k. In addition, we also look at post-war data. There appears to be a strong consensus that the time series properties of stock data differ significantly in the pre- and post-war periods. In particular, predictability seems to be mostly a post-war phenomenon (e.g., Hodrick, 1992; Nelson and

Kim, 1993). Both of the NYSE post-war data sets consist of 480 basic observations (12/1947 to 12/1986), the S&P 500 data set consists of 577 observations (12/1947 to 01/1995).

Our strategy is to construct 95% confidence intervals for the regression parameter β_k and to check whether zero is contained in the intervals or not. We use two-sided symmetric confidence intervals (see end of Section 3) in conjunction with the calibration technique described in Section 4. For the reader interested in the details of the implementation some remarks are in order. They refer to the Description of the Calibration Method from Section 4.

- We used the moving blocks bootstrap with block size $b_{MB} = 100$ to generate the pseudo sequences in step 1.
- To find reasonable block sizes for the subsampling method we estimated the block calibration function $g(\cdot)$ ("how does actual coverage depend on the block size used?"). The so chosen block sizes were between $b = 40$ and $b = 160$, with the great majority of them between $b = 60$ and $b = 120$. The block sizes were larger for the entire data sets than for the post-war data sets.
- A practical issue is the number K of bootstrap samples that we generate in order to estimate the calibration function $h(\cdot)$. We chose $K = 1000$.

The resulting confidence intervals are listed in Table 5 and Table 6.

From the tables we can see that there is no evidence for predictability for horizons of 1, 12, and 24 months, as zero is contained in all corresponding confidence intervals. For the horizon of 36 months, the findings are inconclusive. Of the complete-data intervals, the two NYSE index intervals contain zero, while the S%P 500 interval does not. Of the post-war data intervals, only the equal-weighted NYSE index interval contains zero, the other two do not. For the horizon of 48 months, there appears to be evidence for predictability. None of the intervals contain zero, except for the post-war data set of the equal-weighted NYSE index. In general, the case for predictability seems somewhat stronger for the post-war data.

At this point, it is natural to ask two questions. First, the simulation study in Section 5 suggests some undercoverage of subsampling intervals at long horizons, due to the very strong correlation of the residuals. How much does this evidence take away from the case for predictability that can be made at the four year horizon from the above results? Secondly, we always look at five return horizons simultaneously, namely $k = 1, 12, 24, 36,$ and 48. If

we are more interested in the overall null hypothesis of no predictability rather than in individual hypotheses concerning particular horizons, it seems preferable to derive a test for the joint null hypothesis of $\beta_1 = \beta_{12} = \beta_{24} = \beta_{36} = \beta_{48} = 0$. This avoids the usual pitfalls of multiple testing.

We will deal with both questions in the next section.

7 Additional Looks at Return Regressions

7.1 A Reorganization of the Long-Horizon Regression

Since the compound k-period return is simply the sum of k one-period returns, the numerator of the regression coefficient β_k in equation (2) is the same as

$$Cov[\ln(R_{t+1}) + \ldots + \ln(R_{t+k}),\ (D_t/P_t)]. \tag{14}$$

Under the assumption of stationarity the covariance (14) is identical to

$$Cov[\ln(R_{t+1}),\ (D_t/P_t) + \ldots + (D_{t-k+1}/P_{t-k+1})]. \tag{15}$$

which is the numerator of $\tilde{\beta}_k$ in the following, reorganized regression

$$\ln(R_{t+1}) = \tilde{\alpha}_k + \tilde{\beta}_k[(D_t/P_t) + \ldots + (D_{t-k+1}/P_{t-k+1})] + u_{t+1}. \tag{16}$$

The test $H_0 : \beta_k = 0$ is therefore equivalent to the test $H_0 : \tilde{\beta}_k = 0$. This fact has been recognized by Hodrick (1992), among others. The advantage of the latter test is that, under the null hypothesis, the stochastic behavior of the error terms u_{t+1} in (15) is determined by the behavior of the one-period returns $\ln(R_{t+1})$ only, regardless of the horizon k. Hence, the problem of increasing correlation in the error terms due to an increasing return horizon is eliminated. Remember that, again under the null hypothesis, the error terms $\epsilon_{t+k,k}$ in the regression (2) behave like an MA($k-1$) process, at least under the additional assumption of uncorrelated returns.

Hodrick (1992) carries out this alternative test, using critical values obtained by simulating from a VAR model which imposes the null hypothesis. He still finds evidence for predictability at horizons of one year and beyond. The problem is that the critical values might be too small, since the conditionally normal GARCH innovations of the VAR model tend to underestimate the tails of the true sampling distribution.

To provide an alternative view point we apply the subsampling method. The method of inference about $\tilde{\beta}_k$ is, of course, identical to the method of inference about β_k. Results for the post-war data are reported in Table 7. Notice that all confidence intervals contain zero, and therefore not even at the four year horizon a case for predictability could be made. Results for the complete data were analogous (zero contained in all intervals) and are not reported.

7.2 A Joint Test for Multiple Return Horizons

We now turn to the problem of making joint inference about $\beta = (\beta_1, \beta_{12}, \beta_{24}, \beta_{36}, \beta_{48})$. The null hypothesis of interest is that $\beta = (0, 0, 0, 0, 0)$. The joint estimation is easily done by combining the individual estimates for each horizon into a vector. However, the joint inference is more complicated. Under reasonable conditions, the vector $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_{12}, \hat{\beta}_{24}, \hat{\beta}_{36}, \hat{\beta}_{48})$ will have a limiting normal distribution, centered at β. Obviously, the limiting covariance matrix is not diagonal and can therefore not be estimated by simply combining the individual variance estimates. To make matters worse, the explicit estimation of the 5 by 5 covariance matrix of $\hat{\beta}$ requires along the way the estimation of the 10 by 10 covariance matrix of $(\hat{\alpha}_1, \hat{\beta}_1, \ldots, \hat{\alpha}_{48}, \hat{\beta}_{48})$. This appears not a promising endeavor with sample sizes on the order of 500. Hodrick (1992) runs into this problem when he tries to estimate the limiting covariance matrix by GMM but finds that "simultaneous estimation of the five equations ...results in failure of the GMM matrix to be positive definite". Since in this instance the limiting covariance matrix cannot be estimated, Hodrick is unable to test the null hypothesis of $\beta = (0, 0, 0, 0, 0)$.

Fortunately, the subsampling methodology can be extended to handle multivariate parameters without much difficulty and avoids the problem of explicit estimation of the limiting distribution. The crux of the method is the same as in the univariate parameter case, outlined in Section 3. In the notation of that section, assume now that θ is parameter in $\mathbb{R}^p$, with $p > 1$. Also, $\hat{\theta}_n$ is the estimator of θ based on the entire sample, and $\hat{\theta}_{b,a}$ the estimator based on the block of data $X_a, \ldots, X_{a+b-1}$. We estimate the (multivariate) sampling distribution of $\tau_n(\hat{\theta}_n - \theta)$ by the empirical distribution of the $n - b + 1$ subsample statistics $\tau_b(\hat{\theta}_{b,a} - \hat{\theta}_n)$, $a = 1, \ldots, n - b + 1$. With the help of a norm $\|\cdot\|$ on $\mathbb{R}^p$ we can then find a confidence region for θ quite easily — obvious norm choices are the Euclidean norm or the sup norm. Suppose we want a $1 - \alpha$ confidence region for θ. An asymptotically correct region is given by the collection of all vectors θ' that satisfy

$$\tau_n \left\| \hat{\theta}_n - \theta' \right\| \leq c_{n,\|\cdot\|}(1 - \alpha). \tag{17}$$

Here, $c_{n,\|\cdot\|}(1-\alpha)$ is an $1-\alpha$ quantile of the univariate "normed" subsampling distribution $L_{n,\|\cdot\|}$ having distribution function

$$L_{n,\|\cdot\|}(x) = \frac{1}{n-b+1} \sum_{a=1}^{n-b+1} 1\{\tau_b \left\| \hat{\theta}_{b,a} - \hat{\theta}_n \right\| \leq x\}. \tag{18}$$

See Politis, Romano, and Wolf (1997) for corresponding theoretical results.

While it would be cumbersome to explicitly write down a such-defined confidence region, it is trivial to check whether a specific vector θ' is contained in the region. All we have to do is check condition (17). For our application of stock return regressions, we are obviously interested in the vector $(0,0,0,0,0)$. Also, for this application the proper normalizing constant is again simply the square-root of the sample size, that is, $\tau_b = b^{1/2}$ and $\tau_n = n^{1/2}$.

The problem of choosing the block size b is analogous to the univariate case. We can use the same remedy, the calibration technique described in Section 4. The modifications of the algorithm outlined there should be obvious. Note that to do step 2. the explicit computation of the confidence region in step 1a. is not really needed. All we have to do is check whether $\hat{\theta}$ is contained in the region which, as just pointed out, is an easy matter.

When applying this method to the joint vector β for stock return regression we should be concerned with the magnitudes of the individual coefficients. Note that $\hat{\beta}_k$ naturally will increase with the return horizon k, as we are predicting a k-horizon compounded return. It therefore seems sensible to standardize by dividing by the return horizon. We thus use the following modified Euclidean norm

$$\|(\beta_1, \beta_{12}, \ldots, \beta_{48})\|_{mod} = \sqrt{\beta_1^2 + (\beta_{12}/12)^2 + \ldots + (\beta_{48}/48)^2}\,. \tag{19}$$

The results for the post-war data are reported in Table 8. For all three data sets a block size of $b = 80$ was used. Since the reader might wish some more details rather than simply whether the vector $(0,0,0,0,0)$ is contained in the corresponding confidence region, we decided to give the following information. The observed norm gives the numerical value of $\left\|(\hat{\beta}_1 - 0, \hat{\beta}_{12} - 0, \ldots, \hat{\beta}_{48} - 0)\right\|_{mod}$, with $\|\cdot\|_{mod}$ as defined in (19). The observed P-value reports the percentage of subsample statistics $b^{1/2} \left\|\hat{\beta}_{b,a} - \hat{\beta}_n\right\|_{mod}$ exceeding the scaled observed norm $n^{1/2} \left\|\hat{\beta}_n - 0\right\|_{mod}$. Finally, the cut-off point says how small the observed P-value has to be to be deemed significant at the 5% level by the calibration technique of Section 4. In other words, if the observed P-value is bigger than the cut-off point, then the vector (0,0,0,0,0) is contained in the 95% confidence region.

Note that for all three data sets the observed P-value is substantially bigger than the cut-off point for the 5% level. Hence, for all three data sets $(0,0,0,0,0)$ is contained in the 95% confidence region or, equivalentyly, the null hypothesis of no joint predictability is not rejected. Note that the results for the complete data were identical ($(0,0,0,0,0)$ contained in all three 95% confidence regions) and no closer details are reported.

8 Conclusions

In this article we presented a new statistical tool to make inference in the context of dependent and possibly nonstationary observations, as needed when examining the predictability of stock returns from dividend yields. The gist of the new method, called subsampling, is to recompute the statistic of interest on smaller blocks of the entire data sequence to approximate the sampling distribution of the estimator based on the complete data. This enables us to construct asymptotically correct confidence regions for unknown parameters under very weak conditions.

When comparing the subsampling method with previous approaches for testing the predictability of stock returns, we found it more trustworthy than the VAR approach and Goetzmann and Jorion's (1993) bootstrap on grounds of asymptotic consistency. A simulation study revealed that, for a reasonable data generating mechanism, subsampling appears to have better small sample properties than GMM, which is a valid competitor in terms of asymptotic properties.

We applied the subsampling method to three different data sets, the NYSE equal- and value-weighted indices and the S&P 500 index. We considered complete data sets (starting in 12/1926) as well as post-war data (starting in 12/1947), and included five return horizons ranging between one month and four years. We did not find any evidence for predictability for short and medium horizons, but findings at the four-year horizon appeared significant. However, mild undercoverage of subsampling confidence intervals for long horizons due to very strong dependencies in the residuals, as suggested by our simulation study, and the issue of multiple testing cast some doubt on this evidence.

A reorganization of long-horizon returns, avoiding increasing correlation in the residuals by means of summing dividend yields rather than returns, resulted in insignificant outcomes for all horizons. Moreover, a joint test for all five return horizons also failed to find any evidence. We therefore conclude that no strong case for the predictability of stock returns from dividend yields can be made.

REFERENCES

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–858.

Bekaert, G. and Urias, M. S. (1996). Diversification, integration and emerging market closed-end funds. *Journal of Finance* **3**, 835–869.

Bollerslev, T. (1990). Modeling the coherence on short run nominal exchange rates: a multivariate GARCH model. *Review of Economics and Statistics* **72**, 498–505.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Annals of Statistics* **14**, 1709–1722.

Bühlmann, P. (1994). Blockwise bootstrapped empirical process for stationary sequences. *Annals of Statistics* **22**, 995–1012.

Campbell, J. Y. and Shiller, R. J. (1989). The dividend-price ratio and expectations of future dividends. *Review of Financial Studies* **1**, 195–228.

Campbell, J. Y. and Shiller, R. J. (1988). The dividend ratio model and small sample bias: A Monte Carlo study. *Economics Letters*, **29**, 325–331.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The econometrics of financial markets*, Princeton University Press.

Carlstein, E. (1986). The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *Annals of Statistics* **14**, 1171–1179.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. and Tibshirani, R. J. (1986)). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54–75.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*, Chapman & Hall, New York.

Fama, E. and French, K. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics* **22**, 3–25.

Ferson, W. E. and Foerster S. (1994). Finite sample properties of the generalized method of moments in tests of conditional asset pricing models. *Journal of Financial Economics* **36**, 29–55.

Franke, J. and Härdle W. (1992). On bootstrapping kernel spectral estimates. *Annals of Statistics* **20**, 121–145.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics* **9**, 1218–1228.

Freedman, D. A. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models. *Annals of Statistics* **12**, 827–842.

Goetzmann, W. N. and Jorion, P. (1993). Testing the predictive power of dividend yields. *Journal of Finance* **48**, No. 2, 663–679.

Goetzmann, W. N. and Jorion, P. (1995). A longer look at dividend yields. *Journal of Business* **68**, No. 4.

Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of the American Statistical Association* **83**, 102–110.

Hall, P. (1988). On symmetric bootstrap confidence intervals. *Journal of the Royal Statistical Society* B **50**, No. 1, 35–45.

Hansen, L. and Hodrick, R. (1980). Forward exchange rates as optimal predictors of future spot rates. *Journal of Political Economy*, **88**, 829–853.

Hansen, L. (1982). Large sample properties of generalized method of moments estimation. *Econometrica* **50**, 1029–1054.

Hodrick, R. J. (1992). Dividend yields and expected stock returns: alternative procedures for inference and measurement. *Review of Financial Studies* **5**, No. 3, 357–386.

Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika* **41**, 403–404.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17**, 1217–1241.

Liu, R. Y. (1988) Bootstrap procedures under some non-iid models. *Annals of Statistics* **16**, 1696–1708.

Liu, R. Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the limits of bootstrap*, ed. by LePage and Billard. John Wiley, New York.

Loh, W.-Y. (1987). Calibrating confidence coefficients. Journal of the American Statistical Association, **82**, 155–162.

Lo, A. W. and MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies* **1**, 41–66.

Nelson, C. R. and Kim, M. J. (1993). Predictable stock returns: the role of small sample bias. *Journal of Finance* **48**, No. 2, 641–661.

Noreen, E. W. (1989). *Computer intensive methods for testing hypotheses: An introduction.* John Wiley, New York.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* **22**, 2031–2050.

Politis, D. N. Romano, J. P., and Wolf, M. (1997). Subsampling for heteroskedastic time series. To appear in *Journal of Econometrics.*

Rozeff, M. (1984). Dividend yields are equity risk premium. *Journal of Portfolio Management* **11**, 68–75.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics* **9**, 1187–1195.

Wu, C. F. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics* **14**, 1261–1343.

Table 1: **Parameter estimates for VAR matrix**

This table presents least squares estimates for the VAR matrix A of the following first-order vector-autoregressive model: $Z_{t+1} = AZ_t + u_{t+1}$. Here, Z_t is the joint vector of log return and dvidend yield, $Z_t = [\ln(R_t) - E(\ln(R_t)), D_t/P_t - E(D_t/P_t))]^T$, and u_t is white noise. The estimates are based on monthly post-war data starting in December 1947.

NYSE equal-weighted, 12/1947 to 12/1986
$A = \begin{pmatrix} 0.154 & -0.004 \\ 0.325 & 0.985 \end{pmatrix}$
NYSE value-weighted, 12/1947 to 12/1986
$A = \begin{pmatrix} 0.062 & -0.002 \\ 0.423 & 0.984 \end{pmatrix}$
S&P 500, 12/1947 to 01/1995
$A = \begin{pmatrix} 0.004 & 0.001 \\ 0.422 & 0.984 \end{pmatrix}$

Table 2: **Parameter estimates for GARCH model**

This table presents parameter estimates for the GARCH(1,1) model for the white noise innovation sequence u_t of the VAR. Let $H_t = E_t(u_{t+1}u_{t+1}^T)$ be the conditional covariance matrix with typical element $h_{ij,t}$. Both conditional variances follow an ARMA(1,1) process:

$$h_{ii,t} = \omega_i + \beta_i h_{ii,t-1} + \alpha_i u_{i,t}^2, \quad i = 1, 2.$$

The conditional covariance is determined by a constant correlation coefficient ρ_{12} in the following way: $h_{12,t} = \sqrt{h_{11,t}\, h_{22,t}}\, \rho_{12}$. All parameters are estimated simulatenously via maximum likelihood. The estimates are based on monthly post-war data starting in December 1947.

NYSE equal-weighted, 12/1947 to 12/1986				
Cond. Variance	ω_i	α_i	β_i	ρ_{12}
$h_{11,t}$	0.1624	0.042	0.856	-0.072
$h_{22,t}$	0.0001	0.062	0.791	
NYSE value-weighted, 12/1947 to 12/1986				
Cond. Variance	ω_i	α_i	β_i	ρ_{12}
$h_{11,t}$	0.1006	0.008	0.923	-0.046
$h_{22,t}$	$5.7\ 10^{-5}$	0.014	0.913	
S&P 500, 12/1947 to 01/1995				
Cond. Variance	ω_i	α_i	β_i	ρ_{12}
$h_{11,t}$	2.0188	0.129	0.857	0.128
$h_{22,t}$	0.0001	0.028	0.879	

Table 3: **Estimated coverage probabilities**

This table presents estimated coverage probabilities of nominal 95% confidence intervals. The data generating process is a VAR with GARCH(1,1) innovations. The null hypothesis of no predictability is enforced by setting the first row of the VAR matrix equal to zero. Hence, the true value of β_k is equal to zero for all return horizons k. Two types of confidence intervals are considered, GMM intervals and calibrated symmetric subsampling intervals. The GMM uses the Quadratic Spectral kernel with the automatic bandwidth selection procedure of Andrews (1991). Estimated coverage probabilities are based on 1000 simulations for each scenario.

NYSE equal-weighted, 12/1947 to 12/1986			
Horizon	GMM	Subsampling	Target
$k = 1$	0.93	0.97	0.95
$k = 12$	0.86	0.95	0.95
$k = 24$	0.83	0.94	0.95
$k = 36$	0.82	0.93	0.95
$k = 48$	0.78	0.90	0.95
NYSE value-weighted, 12/1947 to 12/1986			
Horizon	GMM	Subsampling	Target
$k = 1$	0.96	0.97	0.95
$k = 12$	0.87	0.97	0.95
$k = 24$	0.82	0.96	0.95
$k = 36$	0.78	0.92	0.95
$k = 48$	0.76	0.88	0.95
S&P 500, 12/1947 to 01/1995			
Horizon	GMM	Subsampling	Target
$k = 1$	0.94	0.97	0.95
$k = 12$	0.89	0.96	0.95
$k = 24$	0.84	0.95	0.95
$k = 36$	0.81	0.93	0.95
$k = 48$	0.79	0.91	0.95

Table 4: **VAR innovation 0.95 quantile – GARCH model vs. Observed statistic** This table compares the observed 0.95 quantile of the estimated VAR innovations with the sampling distribution of the corresponding GARCH(1,1) model. The GARCH(1,1) model was obtained via maximum likelihood from the estimated innovations. The sampling distribution is characterized by the 0.01 quantile, the mean, the median, and the 0.99 quantile. The 2-sided P-value tests the null hypothesis that the GARCH(1,1) model gave rise to the fitted innovations.

NYSE equal-weighted, 12/1947 to 12/1986						
	0.01 quantile	Mean	Median	0.99 quantile	Observed	2-sided P-value
Log returnd	1.782	2.074	2.069	2.363	5.554	0
Dividend yield	0.044	0.051	0.051	0.059	0.339	0
NYSE value-weighted, 12/1947 to 12/1986						
	0.01 quantile	Mean	Median	0.99 quantile	Observed	2-sided P-value
Log returnd	1.732	1.983	1.983	2.254	7.764	0
Dividend yield	0.040	0.046	0.046	0.052	0.355	0
S%P 500, 12/1947 to 01/1995						
	0.01 quantile	Mean	Median	0.99 quantile	Observed	2-sided P-value
Log return	11.564	18.482	17.112	40.275	6.105	0
Dividend yield	0.058	0.067	0.067	0.076	0.331	0

Table 5: **95% confidence intervals for β_k, Complete data**

This table presents 95% confidence intervals for the return regression coefficient β_k, together with the estimated coeffficient $\hat{\beta}_k$. We use monthly data, and various return horizons k are considered. The confidence intervals are calibrated symmetric subsampling intervals. They are based on the complete data, starting in December 1926.

NYSE equal-weighted, 12/1926 to 12/1986		
Horizon	$\hat{\beta}_k$	95% CI
$k = 1$	0.18	[-0.50, 0.86]
$k = 12$	4.08	[-4.32, 12.49]
$k = 24$	9.79	[-3.37, 22.96]
$k = 36$	13.29	[-1.04, 27.62]
$k = 48$	16.16	[2.51, 29.81]
NYSE value-weighted, 12/1926 to 12/1986		
Horizon	$\hat{\beta}_k$	95% CI
$k = 1$	0.27	[-0.26, 0.79]
$k = 12$	4.53	[-0.77, 9.82]
$k = 24$	8.96	[-3.71, 21.63]
$k = 36$	11.95	[-0.67, 24.57]
$k = 48$	15.18	[9.43, 20.94]
S&P 500, 12/1926 to 01/1995		
Horizon	$\hat{\beta}_k$	95% CI
$k = 1$	0.29	[-0.13, 0.72]
$k = 12$	3.46	[-2.25, 9.18]
$k = 24$	6.04	[-1.68, 13.75]
$k = 36$	8.44	[2.85, 14.02]
$k = 48$	11.13	[6.65, 15.61]

Table 6: **95% confidence intervals for β_k, Post-war data**

This table presents 95% confidence intervals for the return regression coefficient β_k, together with the estimated coeffficient $\hat{\beta}_k$. We use monthly data, and various return horizons k are considered. The confidence intervals are calibrated symmetric subsampling intervals. They are based on the post-war data, starting in December 1947.

NYSE equal-weighted, 12/1947 to 12/1986		
Horizon	$\hat{\beta}_k$	95% CI
$k = 1$	0.28	[-0.98, 1.54]
$k = 12$	4.54	[-11.99, 21.07]
$k = 24$	8.70	[-24.55, 41.94]
$k = 36$	11.32	[-17.90, 40.55]
$k = 48$	13.24	[-10.55, 37.02]
NYSE value-weighted, 12/1947 to 12/1986		
Horizon	$\hat{\beta}_k$	95% CI
$k = 1$	0.41	[-0.45, 1.28]
$k = 12$	5.67	[-5.22, 16.57]
$k = 24$	10.40	[-5.54, 26.34]
$k = 36$	13.81	[0.08, 27.53]
$k = 48$	17.37	[7.50, 27.24]
S&P 500, 12/1947 to 01/1995		
Horizon	$\hat{\beta}_k$	95% CI
$k = 1$	0.42	[-0.72, 1.57]
$k = 12$	5.38	[-3.99, 14.76]
$k = 24$	8.88	[-1.02, 18.78]
$k = 36$	11.82	[3.43, 20.21]
$k = 48$	14.82	[6.69, 22.94]

Table 7: **95% confidence intervals for $\tilde{\beta}_k$, Post-war data**

This table presents 95% confidence intervals for the return regression coefficient $\tilde{\beta}_k$ of the reorganized regression (16) which avoids additional correlation in the residuals for long return horizons. Also, the estimated coefficient $\hat{\tilde{\beta}}_k$ is presented. We use monthly data, and various return horizons k are considered. The confidence intervals are calibrated symmetric subsampling intervals. They are based on the post-war data, starting in December 1947.

NYSE equal-weighted, 12/1947 to 12/1986		
Horizon	$\hat{\tilde{\beta}}_k$	95% CI
$k = 1$	0.281	[-0.98, 1.54]
$k = 12$	0.035	[-0.19, 0.27]
$k = 24$	0.019	[-0.24, 0.28]
$k = 36$	0.011	[-0.13, 0.15]
$k = 48$	0.007	[-0.07, 0.08]
NYSE value-weighted, 12/1947 to 12/1986		
Horizon	$\hat{\tilde{\beta}}_k$	95% CI
$k = 1$	0.412	[-0.45, 1.28]
$k = 12$	0.043	[-0.15, 0.24]
$k = 24$	0.022	[-0.10, 0.15]
$k = 36$	0.013	[-0.13, 0.16]
$k = 48$	0.009	[-0.04, 0.05]
S&P 500, 12/1947 to 01/1995		
Horizon	$\hat{\tilde{\beta}}_k$	95% CI
$k = 1$	0.425	[-0.72, 1.57]
$k = 12$	0.040	[-0.10, 0.18]
$k = 24$	0.018	[-0.05, 0.08]
$k = 36$	0.011	[-0.04, 0.06]
$k = 48$	0.007	[-0.01, 0.03]

Table 8: **Joint test for $\beta_1 = \beta_1 2 = \ldots = \beta_{48} = 0$, Post-war data**
This table presents results for the joint test of all individual regression coefficients being equal to zero. The observed norm gives the numerical value of $\left\|(\hat{\beta}_1 - 0, \hat{\beta}_{12} - 0, \ldots, \hat{\beta}_{48} - 0)\right\|_{mod}$, with $\|\cdot\|_{mod}$ as defined in (19). The observed P-value reports the percentage of subsample statistics $b^{1/2}\left\|\hat{\beta}_{b,a} - \hat{\beta}_n\right\|_{mod}$ exceeding the scaled observed norm $n^{1/2}\left\|\hat{\beta}_n - 0\right\|_{mod}$. Finally, the cut-off point says how small the observed P-value has to be to be deemed significant at the 5% level by the calibration technique of Section 4. In other words, if the observed P-value is bigger than the cut-off point, then the vector (0,0,0,0,0) is contained in the 95% confidence region. The results are based on the post-war data, starting in December 1947.

NYSE equal-weighted, 12/1947 to 12/1986		
Observed Norm	Observed P-value	Cut-off point for 0.05 test
0.727	0.518	0.027
NYSE value-weighted, 12/1947 to 12/1986		
Observed Norm	Observed P-value	Cut-off point for 0.05 test
0.927	0.385	0.060
S&P 500, 12/1947 to 01/1995		
Observed Norm	Observed P-value	Cut-off point for 0.05 test
0.850	0.268	0.081