SECULAR MEAN REVERSION AND LONG-RUN PREDICTABILITY OF THE STOCK MARKET

Valeriy Zakamulin
School of Business and Law, University of Agder

ABSTRACT

The empirical financial literature reports evidence of mean reversion in stock prices and the absence of out-of-sample return predictability over horizons shorter than 10 years. Anecdotal evidence suggests the presence of mean reversion in stock prices and return predictability over horizons longer than 10 years, but thus far, there is no empirical evidence confirming such anecdotal evidence. The goal of this paper is to fill this gap in the literature. Specifically, using 141 years of data, this paper begins by performing formal tests of the random walk hypothesis in the prices of the real S&P Composite Index over increasing time horizons of up to 40 years. Although our results cannot support the conventional wisdom that the stock market is safer for long-term investors, our findings speak in favor of the mean reversion hypothesis. In particular, we find statistically significant in-sample evidence that past 15-17 year returns are able to predict the future 15-17 year returns. This finding is robust to the choice of data source, deflator, and test statistic. The paper continues by investigating the out-of-sample performance of long-horizon return forecasting based on the mean-reverting model. These latter tests demonstrate that the forecast accuracy provided by the mean-reverting model is statistically significantly better than the forecast accuracy provided by the naive historical-mean model. Moreover, we show that the predictive ability of the mean-reverting model is economically significant and translates into substantial performance gains.

Keywords: bootstrap simulation, long run, mean reversion, random walk, predictability, stock returns

JEL classification number: C12, C14, C22, G12, G14, G17

I. INTRODUCTION

Until the late 1980s, there was widespread agreement in the academic community that stock prices follow a random walk. Indeed, a large body of empirical literature seemed to support this perspective (see Fama, 1970 and Leroy, 1982 for surveys). The efficient market hypothesis is

Correspondence to: Valeriy Zakamulin, School of Business and Law, University of Agder, Service Box 422, 4604 Kristiansand, Norway. E-mail: Valeri.Zakamouline@uia.no. The author is grateful to the following individuals for their helpful comments and suggestions on previous drafts of this paper: two anonymous referees; Brian Berry; Ed Easterling; Jochen Jungeilges; Ole Gjølberg; Steen Koekebakker; and the participants at the 2011 Computational and Financial Econometrics Conference (London, UK), the 2013 Fagkonferansen i Bedriftsøkonomiske Emner (FIBE Conference, Bergen, Norway), and the 2013 Forecasting Financial Markets Conference (Hannover, Germany). Any remaining errors in the manuscript are the author’s responsibility.
strongly associated with the idea of a random walk in stock prices and, loosely, states that stock returns are unpredictable. However, during the late 1980s, a series of papers appeared in which the authors challenged the random walk hypothesis (see, for example, Summers, 1986, Fama and French, 1988b, Lo and MacKinlay, 1988, and Poterba and Summers, 1988). In particular, these authors considered the time series properties of stock returns over increasing time horizons of up to 10 years and found indications of mean reversion and return predictability. For example, Fama and French (1988b) discovered a substantial negative autocorrelation in returns over periods of 3-5 years and concluded that the past 3-5 year returns are able to predict the future 3-5 year returns. Poterba and Summers (1988) found that stock returns exhibit positive and statistically significant autocorrelation in returns over periods shorter than one year and negative, although not statistically significant at conventional levels (1 percent or 5 percent), autocorrelations over longer periods.

However, the conclusions reached in these earlier papers have been strongly criticized on statistical grounds. For example, Kim et al. (1991) demonstrated that due to small-sample bias, the statistical significance of the test statistics in Fama and French (1988b) and Poterba and Summers (1988) was overstated and that the future 3-5 year returns could not be predicted on the basis of the past 3-5 year returns. Similarly, Richardson and Stock (1989) and Richardson (1993) showed that correcting for small-sample bias may reverse the results obtained by Fama and French (1988b) and Poterba and Summers (1988). Apparently, the statistical power of earlier tests was insufficient to reject the random walk hypothesis. Jegadeesh (1991) suggested a new, more powerful test and detected statistically significant evidence of mean reversion in stock prices (over periods of 4-8 years). In addition, Jegadeesh found evidence of mean reversion not only for the US stock market but also for the UK stock market. Subsequently, based on a panel approach, Balvers et al. (2000) found statistically significant evidence of mean-reverting behavior (over periods of 3-3.5 years) in many international stock indices. Thus, mean reversion in stock prices seems to be an international phenomenon. Using the same technique as in Balvers et al. (2000), Gropp (2003) and Gropp (2004) found statistically significant evidence of mean reversion in the prices of portfolios of small-cap stocks (over periods of 3.5 years) and industry-sorted portfolios (over periods of 4.5-8 years). Moreover, Balvers et al. (2000), Gropp (2003), and Gropp (2004) showed that parametric contrarian investment strategies that exploit mean reversion outperform buy-and-hold and standard contrarian strategies. This provides further support for the mean reversion findings in these papers.

Thus, recent research appears to have provided evidence of mean reversion in the prices of some stock portfolios over periods of 3-8 years. In contrast, the predictability of stock returns remains a source of heated debate within the academic community. Earlier papers that demonstrated the existence of in-sample stock return predictability include Fama (1981), Campbell (1987), Fama and French (1988b), Fama and French (1988a), Campbell and Shiller (1988), and Fama and French (1989). Again, the conclusions reached in these earlier papers have been strongly criticized on statistical grounds. For example, Richardson and Stock (1989)

---

1 Mean reversion is an ambiguous concept and exists in several different forms. Typically, the concept of mean reversion can be expressed by the common investment wisdom that ‘over time, markets tend to return to the mean’. For example, when stocks move too far in one direction, this movement will eventually be reversed. Another type of mean reversion, which is studied in this paper, implies that reversion is much more than simply returning to the mean. In reality, the movement is far greater. This type of mean reversion incorporates further common investment wisdom, namely, that ‘an excess in one direction will lead to an excess in the opposite direction’. That is, when stocks move too far in one direction, they will not merely return to the mean but overshoot in the opposite direction. For example, a period of above-average returns tends to be followed by a period of below-average returns and vice versa. Throughout the paper, the term ‘period’ is used to denote the period of mean reversion. The term ‘horizon’ is used primarily to denote the average length of a complete cycle of reversion, which consists of two periods: a period of higher than average returns and a period of lower than average returns (or vice versa).
and Nelson and Kim (1993) highlighted the small-sample bias problem, whereas Cavanagh et al. (1995), Stambaugh (1999), and Lanne (2002) noted the neglect of the potential for a near unit root problem. Responding to such critiques, Lewellen (2004), Torous et al. (2004), Rapach and Wohar (2005), and Campbell and Yogo (2006) developed new tests that are free of the flaws discovered in the earlier tests and again found some evidence of in-sample predictability. Yet, Bossaerts and Hillion (1999), Goyal and Welch (2003), and Welch and Goyal (2008) demonstrated that, despite evidence of in-sample predictability, the predictive models have no out-of-sample forecasting power. These authors therefore argued that in-sample predictability appears as a result of data mining. It should be noted, however, that the longest forecast horizon considered in all these tests was 10 years. Consequently, the results of these tests imply that the predictive models fail to demonstrate statistically significant predictive ability over short- and medium-term horizons.

To the best of our knowledge, no study to date has tested the random walk hypothesis in stock prices over horizons longer than 10 years. Yet, anecdotal evidence suggests the presence of mean reversion in stock prices over very long horizons. Likely the best known evidence is presented by Siegel (2002) in his famous book “Stocks for the Long Run”. In particular, using a historical sample that covers nearly 200 years, Siegel computed the standard deviation of average real annual returns on a broad US stock market index over increasing horizons of up to 30 years. Siegel found that the standard deviation declines far more rapidly than predicted by the random walk hypothesis. This led many to conclude that stocks are less risky over the long run. However, thus far, there have been no studies of whether the decline in the standard deviation over very long horizons is statistically significant.

Another piece of well-known anecdotal evidence, explicitly related to mean reversion in stock prices over very long horizons, suggests the existence of long-lasting alternating periods of bull and bear markets. These long-lasting bull and bear markets are often dubbed ‘secular’ bull and bear markets. Alexander (2000), Easterling (2005), Rogers (2005), Katsenelson (2007), and Hirsch (2012), among others, analyzed the dynamics of the real S&P Composite Index since 1870 and found indications of secular stock market trends that lasted from 5 to 25 years, with an average duration of approximately 15 years. Motivated by the seeming regularity in the reversion of secular trends, some authors made quite successful forecasts of the long-run US stock market outlook. For example, Alexander (2000) predicted that during the period from 2000 to 2020 the stock market would not beat the money market. Thus far, this forecast has proven true. This anecdotal evidence suggests, among other things, that a price change over a given long-run period may be able to predict the price change over the subsequent long-run period. This idea serves as motivation to re-examine the predictive performance of the model introduced by Fama and French (1988b). Although Kim et al. (1991) demonstrated that this model has no predictive power on increasing horizons of up to 10 years, to the best of our knowledge, no study has tested this model for horizons longer than 10 years. The aim of this paper is to fill these gaps in the literature regarding stock market dynamics over very long horizons.

The first main contribution of this paper is to provide, for the first time, statistically significant evidence against the random walk hypothesis over periods longer than 10 years. Although our results cannot fully support the anecdotal evidence stating that the stock market is safer for long-term investors, our findings do suggest mean reversion in stock prices over periods of 15-17 years. In particular, using our full sample, we find statistically significant evidence that a given change in prices over 15-17 years tends to be reversed over the next 15-17 years by a predictable change in the opposite direction. This implies the existence of in-sample long-horizon predictability. Since the conventional wisdom holds that in-sample evidence of stock return predictability might be a result of data mining, we investigate the performance of out-of-sample long-horizon return forecasts. In addition to the mean-reverting model, we investigate
the out-of-sample forecast accuracy of a few other competing models that employ, as a predictor of long-horizon returns, the cyclically adjusted price-to-earnings ratio, the price-to-dividends ratio, and the long-term bond yield.

The second main contribution of this paper is to demonstrate that the out-of-sample long-horizon forecasts provided by the mean-reverting model and the models that employ the price-to-earnings and price-to-dividends ratios are statistically significantly better than the forecast provided by the historical-mean model. It is worth emphasizing that Welch and Goyal (2008) also used the price-to-earnings and price-to-dividends ratios in their study and found that these models have no predictive ability over forecast horizons of up to 5 years. Our results therefore demonstrate that these models do have predictive ability, but over forecast horizons longer than 10 years. We also demonstrate that the advantages of the models, which demonstrate their predictive ability, translate into significant performance gains. For example, we estimate that risk-averse investors would be willing to pay from 30 to 77 basis points in fees per year to switch from the historical-mean model to a model with superior forecast accuracy. Moreover, our tests suggest that in the recent past, the out-of-sample forecast accuracy provided by the mean-reverting model was substantially better than that provided by the competing models. In addition, we find that the mean-reverting model delivers the highest performance gains when investors have to make long-term allocation decisions.

Overall, our paper contributes to the finance literature in several respects. First, it expands our understanding of stock market dynamics over longer horizons. A major problem with studying stock market behavior over longer horizons is limited data. For example, using data that span 141 years to study the distribution of stock returns over a period of 20 years, researchers have only seven truly independent observations. It is widely believed that tests based on such small sample sizes are bound to have low statistical power. Our paper demonstrates that, using overlapping blocks of data and randomization methods, one can dramatically increase statistical power of tests. Second, our study reaches several important conclusions for long-term investors. The gain in test power delivered by our approach allows us to find weak support for the conventional wisdom that the stock market is safer for long-term investors and strong support for the existence of secular mean reversion in the stock market. Third, the evidence of stock market return predictability over horizons shorter than 10 years is conflicting and, essentially, negative. In other words, investors cannot rely on short- and medium-term stock market return forecasts. This paper convincingly demonstrates that secular mean reversion can be used to forecast long-term stock market returns. Although the improvement in the forecast accuracy provided by the mean-reverting model, relative to that delivered by other alternative models, is not especially impressive when judged by the mean squared error, our predictive model generates meaningful utility gains for mean-variance investors. Thus, the results reported in this paper allow investors to make better long-term allocation decisions.

The remainder of the paper is organized as follows. Section II presents the data for our study, namely, the returns on the real Standard and Poor’s Composite Stock Price Index over the period from 1871 to 2011. In Section III, we perform the tests of the random walk hypothesis using the S&P Composite Index. In Section IV, we study the out-of-sample predictability of multi-year returns on the S&P Composite Index. Finally, Section V summarizes and concludes the paper.

II. THE DATA

The data for the study in this paper are the annual log real returns on a broad US stock market index for the period from 1871 to 2011. The returns are adjusted for dividends and computed using the real (i.e., inflation-adjusted) Standard and Poor’s Composite Stock Price Index data.
and corresponding dividend data. The inflation adjustment is done using the US Consumer Price Index (CPI). All the data are provided by Robert Shiller. The Standard and Poor's Composite Stock Price index is a value-weighted stock index. The index for the period from 1871 to 1925 is constructed using the Cowles Commission Common Stock Index series. From 1926 to the present, the index data come from various reports of Standard and Poor's. From 1957, this index is identical to the Standard and Poor's 500 Index, which is intended to be a representative sample of leading companies in leading industries in the US economy. Stocks in the index are chosen for market size, liquidity, and industry group representation. For further details on the construction of the index and its dividend series, see Shiller (1989), Chapter 26. Formally, let \((p_0, p_1, \ldots, p_n)\) be observations of the natural log of an inflation-adjusted stock index price over \(n + 1\) years. Denote the one-year log return during year \(t\), \(1 \leq t \leq n\), by

\[ r_t = \log(p_t) - \log(p_{t-1}). \]

The resulting sample of \(n\) return observations is \((r_1, r_2, \ldots, r_n)\). The probability distribution of \(r_t\) is unknown, yet it is well-documented that stock returns are non-normal and heteroskedastic.

To check the robustness of the findings, in particular, to determine whether the results of testing the random walk hypothesis depend on a specific historical period, we divide the total sample period from 1871 to 2011 (141 annual observations) into two equal, overlapping sub-samples; the first is from 1871 to 1956 and the second is from 1926 to 2011. Both of these sub-samples cover a span of 86 years. Table 1 presents the descriptive statistics for the annual stock index returns, \(r_t\), for the full sample and both sub-samples. Table 2 reports the results of the \(t\)-test on the difference in mean returns and \(F\)-test on the difference in standard deviations between the first and the second sub-sample. The descriptive statistics and the results of the tests suggest that the mean and variance of returns on the index were generally stable during the full sample. Specifically, using a \(t\)-test for equal means, we cannot reject the hypothesis that the mean returns in the two sub-samples are alike. Similarly, using an \(F\)-test for equal variances, we cannot reject the hypothesis that the variances of the two sub-samples are alike. All series exhibit negative skewness and positive excess kurtosis, which indicates a deviation from normality. Observe further that the return series during the full sample period exhibits a statistically significant negative autocorrelation at lag 2 (at the 5% level). There are no other indications of serial dependence in the return series.

III. TESTING THE RANDOM WALK HYPOTHESIS

III.1 Methodology

One of the main questions we seek to study in this paper is whether the log of the real S&P Composite Stock Price Index follows a random walk. To answer this question, we perform two well-known tests. The first test is based on the examination of the first-order autocorrelation function of \(k\)-year returns. This test has been used by, for example, Fama and French (1988b),

See http://www.econ.yale.edu/~shiller/data.htm. The real dividend-adjusted annual return series on the index are readily available in the file chapt26.xls. Robert Shiller stopped maintaining his database in 2012.

The reasons for using overlapping sub-samples are as follows. First, to perform statistical tests on the presence of long-run mean reversion, we need longer time series. Second, the starting point of our second sub-sample coincides with the starting point of the database of historical stock market data provided by the Center for Research in Security Prices. Therefore, the data on the stock market returns over the second sub-sample is much more accurate than that over the first sub-sample.

© 2016 Board of Trustees of the Bulletin of Economic Research and John Wiley & Sons Ltd
TABLE 1
Descriptive statistics of the annual log real returns on the Standard and Poor’s Composite Stock Price Index

| Sample period | 1871-2011 | 1871-1956 | 1926-2011 |
|---------------|-----------|-----------|-----------|
| Mean, %       | 6.28      | 6.91      | 6.24      |
| Std. dev. %   | 17.14     | 17.76     | 18.77     |
| Skewness      | -0.57     | -0.48     | -0.59     |
| Kurtosis      | 3.41      | 3.32      | 3.24      |
| $\rho_1$      | 0.02      | 0.04      | 0.04      |
| $\rho_2$      | **-0.19** | -0.20     | -0.18     |
| $\rho_3$      | 0.09      | 0.07      | 0.02      |
| $\rho_4$      | -0.08     | -0.18     | -0.14     |
| $\rho_5$      | -0.11     | -0.10     | -0.07     |
| $\rho_6$      | 0.10      | 0.12      | 0.11      |
| $\rho_7$      | 0.10      | 0.06      | 0.16      |
| $\rho_8$      | -0.08     | -0.15     | -0.02     |
| $\rho_9$      | -0.06     | -0.04     | 0.04      |
| $\rho_{10}$   | 0.02      | 0.06      | 0.06      |
| $\rho_{11}$   | 0.02      | 0.06      | -0.07     |
| $\rho_{12}$   | -0.08     | -0.04     | -0.10     |
| $\rho_{13}$   | -0.09     | -0.15     | -0.19     |
| $\rho_{14}$   | 0.03      | 0.03      | -0.14     |
| $\rho_{15}$   | -0.09     | -0.07     | -0.02     |
| $\rho_{16}$   | -0.09     | -0.09     | 0.06      |
| $\rho_{17}$   | 0.06      | 0.16      | -0.02     |
| $\rho_{18}$   | -0.08     | -0.06     | -0.13     |
| $\rho_{19}$   | -0.17     | -0.11     | -0.21     |
| $\rho_{20}$   | 0.06      | 0.09      | -0.07     |

Note: $\rho_k$ denotes the autocorrelation between $r_t$ and $r_{t+k}$. For each $\rho_k$, we test the following hypothesis $H_0: \rho_k = 0$. Bold text indicates values that are statistically significant at the 5% level.

TABLE 2
Results of the t-test on the difference in mean returns and F-test on the difference in standard deviations between the first and the second sub-sample

|                          | Test statistic | P-value |
|--------------------------|----------------|---------|
| t-test on difference in mean returns | 0.24           | 0.81    |
| F-test on difference in standard deviations | 0.89           | 0.61    |

Fama and French (1989), and Fama (1990) and is based on the computation of the following test statistic

$$AC1(k) = \frac{Cov(r_{i,t+k}, r_{i-k,i})}{\sqrt{Var(r_{i,t+k})Var(r_{i-k,i})}},$$

(1)

where $r_{i,j}$ is the compounded return from year $i$ to year $j$, $r_{i,j} = p_j - p_i$, $Cov(\cdot, \cdot)$ and $Var(\cdot)$ denote the covariance and variance, respectively, and $AC1(k)$ represents the first-order autocorrelation function of $k$-year returns. The second test is based on the examination of the variance.

© 2016 Board of Trustees of the Bulletin of Economic Research and John Wiley & Sons Ltd
ratio. This test is very popular and has been used by Cochrane (1988), Lo and MacKinlay (1988), Poterba and Summers (1988), and many others. The test is based on the computation of the following test statistic

\[ VR(k) = \frac{Var(r_{t,t+k})}{k \times Var(r_t)}. \]  

Both of the tests are motivated by the notion that if the stock returns are independent and identically distributed, then the first-order autocorrelation function is zero and the variance ratio is unity irrespective of the number of years, \( k \). In other words, absent serial dependence in the data, the variance of \( k \)-year returns equals \( k \) times the variance of one-year returns and there is no correlation between two successive non-overlapping \( k \)-year returns. The null hypothesis of a random walk is rejected if the first-order autocorrelation is significantly different from zero or the variance ratio is significantly different from unity.

We wish to compute the variance ratio, \( VR(k) \), for return horizons, \( k \), from 20 to 40 years and the first-order autocorrelation, AC1\( (k) \), for periods from 10 to 20 years (note that in the latter case, we also study serial dependence in the data over time horizons from 20 to 40 years). The fundamental problem with these computations is that we have only a few non-overlapping intervals of length 20-40 years. Therefore, in the computations of the two test statistics, we employ overlapping intervals (rolling \( k \)-year periods). To compute AC1\( (k) \), we regress \( k \)-year returns \( r_{t,t+k} \) on lagged \( k \)-year returns \( r_{t-k,t} \). That is, we run the following regression

\[ r_{t,t+k} = a(k) + b(k) r_{t-k,t} + \epsilon_{t,t+k}. \]  

Observe that the slopes of the regression, \( b(k) \), \( k \in [10, 20] \), are the estimated autocorrelations of \( k \)-year returns, AC1\( (k) \). The variance of \( k \)-year returns is computed as

\[ Var(r_{t,t+k}) = E[(r_{t,t+k} - E[r_{t,t+k}])^2]. \]

The use of overlapping returns leads to some potentially very serious econometric issues that are commonly termed ‘small-sample bias’. In particular, when estimating regression (3), there are two econometric problems. First, the estimates of the slope coefficients are biased. The sources of this bias in the estimation of autocorrelation are described in detail by Orcutt and Irwin (1948) and Marriott and Pope (1954). Specifically, these authors demonstrated that an estimate of autocorrelation obtained using overlapping blocks of data is downward biased. Therefore, the estimates must be corrected for this bias. The second problem is that the standard errors of estimations using overlapping blocks of data are also downward biased; see, for example, Nelson and Kim (1993). Both biases work in the direction of making the values of the \( t \)-statistic too large such that standard inference may indicate dependence in the return series even if none is present.

Similarly, the estimate of the variance of multi-year returns, \( Var(r_{t,t+k}) \), is downward biased when one uses overlapping blocks of data.\(^5\) As an immediate consequence, the estimate of the

\(^4\)Specifically, when returns are independent, using overlapping blocks of data produces a negative value of the estimated slope coefficient in regression (3). In addition, the standard error of estimation of the slope coefficient using overlapping blocks of data is downward biased. That is, the estimated standard error is smaller than the true value. The greater the overlap is, the more negative the slope coefficient and the smaller the estimated standard error. As a result, the values of the \( t \)-statistic may falsely indicate the presence of dependence in the return series when none is present.

\(^5\)Note that this is also related to the second econometric problem in the estimation of regression (3). That is, the standard errors of estimation of slope coefficients using overlapping blocks of data are downward biased because the estimates of the variance using overlapping blocks of data are downward biased. For the sake of motivation, consider what happens to the estimate of \( Var(r_{t,t+k}) \) when \( k \rightarrow n \). Obviously, in the limit, when the length \( k \) converges to the sample length, there is only one available block of data to estimate \( Var(r_{t,t+k}) \). Therefore, regardless of the nature of the data-generating process, \( Var(r_{t,t+k}) \rightarrow 0 \) as \( k \rightarrow n \).
variance ratio $VR(k)$ also becomes downward biased. Therefore, the estimates of $VR(k)$ must be corrected for this bias. In addition, since the estimate of $VR(k)$ is a random variable, for the purpose of statistical inference, we need to know the probability distribution of $VR(k)$. This is necessary to be able to estimate standard errors and confidence intervals for $VR(k)$. This is also necessary for performing hypothesis tests concerning the value of $VR(k)$.

When the nature of the data-generating process is unknown, it is generally not possible to address the econometric problems described above. However, in the context of the null hypothesis, our goal is primarily to test whether stock returns are distributed independently of their ordering in time. Because, under the null, there is no dependence in return series, to estimate the significance level and perform the bias correction of the test statistics, we closely follow Kim et al. (1991) and Nelson and Kim (1993), who employed the randomization method. The randomization method was introduced by Fisher (1935) and provides a very general and robust approach for computing the probability of obtaining some specific value for an estimator under the null hypothesis of no dependence. We refer interested readers to Noreen (1989) and Manly (1997) for extensive discussions of the randomization tests. In essence, randomization consists of reshuffling the data to destroy any dependence and then recalculating the test statistics for each reshuffling to estimate its distribution under the null hypothesis of no dependence. The great advantage of the randomization method is that it is very simple and requires no assumptions about the actual distribution of stock returns.

Specifically, consider the estimation of the significance level and the bias correction of the estimate of the autocorrelation of $k$-year returns, $AC1(k)$. First, we run regression (3) using the original series $(r_1, r_2, \ldots, r_n)$ to obtain the actual historical estimates of $AC1(k)$. Then, we randomize the original series to obtain a permutation $(r_1^*, r_2^*, \ldots, r_n^*)$. This is repeated 10,000 times, each time running regression (3) and obtaining an estimate of $AC1^*(k)$. In this manner, we estimate the sampling distribution of $AC1(k)$ under the null hypothesis. Finally, to estimate the significance level for some particular $k$, we count how many times the computed value of $AC1^*(k)$ after randomization falls below the value of the actual historical estimate of $AC1(k)$. In other words, under the null hypothesis, we compute the probability of obtaining a more extreme value for the autocorrelation of $k$-year returns than the actual historical estimate. Note that in this manner, we compute the p-values of a one-tailed test. The estimation bias is defined as the difference between the expected and true value of $AC1^*(k)$. Since the true value is zero under the null hypothesis, the bias correction is achieved by subtracting the expected value of $AC1^*(k)$ from the actual historical estimate of $AC1(k)$. That is, the bias-adjusted values of the first-order autocorrelation of $k$-year returns are computed as $AC1(k) - E[AC1^*(k)]$.

The estimation of the significance level and the bias correction of the estimate of the variance ratio $VR(k)$ is performed in a similar manner. First, we use the original series to obtain the actual historical estimates of $VR(k)$. Then, we randomize the series and compute $VR^*(k)$ to obtain the sample distribution under the null hypothesis. Finally, to estimate the significance level for some particular $k$, we count how many times the computed value of $VR^*(k)$ after randomization falls below the value of the actual historical estimate of $VR(k)$. The estimation bias in this case is given by $E[VR^*(k)] - 1$ since the true value is unity under the null hypothesis. Finally, the bias-adjusted values of the variance ratio are computed as $VR(k) - E[VR^*(k)] + 1$.

There is ample evidence that the series of stock returns is heteroskedastic; see, for example, Officer (1973) and Schwert (1989). In particular, many researchers document that the variance of stock returns is not constant, but time-varying. To determine whether a change in the variance of returns might affect the sampling distribution of a test statistic, we closely follow Kim et al. (1991) and Nelson and Kim (1993) and use stratified randomization. In the stratified
The sample first-order autocorrelations (top panel) and variance ratios (bottom panel) for the $k$-year log real returns on the Standard and Poor’s Composite Stock Price Index. Neither the first-order autocorrelations nor the variance ratios are adjusted for estimation bias.

randomization method, the total sample (or a sub-sample) is divided into several separate bins (urns) and the randomization is performed within each bin. Such a stratified randomization allows us to determine whether the sampling distribution of a test statistic is sensitive to the particular pattern of heteroskedasticity that occurred historically.

© 2016 Board of Trustees of the Bulletin of Economic Research and John Wiley & Sons Ltd
III.2 Empirical Results

Figure 1 plots the sample first-order autocorrelations and variance ratios of the $k$-year returns on the Standard and Poor’s Composite Stock Price Index. The first-order autocorrelations and variance ratios are computed according to formulas (1) and (2), respectively, using overlapping blocks of data. In the full sample and both sub-samples, the first-order autocorrelations and variance ratios generally decline with increasing $k$. The indications that the null hypothesis should be rejected for very long horizons are stronger (i.e., the declines in the first-order autocorrelations and variance ratios are larger) for the second sub-sample (1926 to 2011) than for the full sample or the first sub-sample (1871 to 1956).

Recall, however, that the estimates of both the first-order autocorrelations and the variance ratios presented in Figure 1 are downward biased. As a matter of fact, under the null hypothesis of no serial dependence in the return series, we would expect to observe declining first-order autocorrelations and variance ratios with increasing $k$. To determine whether the observed declines are statistically significantly different from the expected declines under the null hypothesis, and to correct for estimation bias under the null, we perform the randomization method with and without the stratification. These results are reported in Tables 3 and 4, which report the estimates for the bias-adjusted first-order autocorrelations and variance ratios, respectively, with corresponding p-values. The estimates are based on 10,000 reshuffles and computed using different numbers of bins in the stratification. The number of bins varies from 1 (no stratification) to 5.

Without the stratification (that is, when the number of bins equals one), both test statistics suggest that the return series over the full sample period (1871 to 2011) and the second sub-sample (1926 to 2011) exhibit clear evidence against the random walk on horizons of approximately 30-40 years. In particular, for the full sample, the first-order autocorrelation values are statistically significantly negative at the 5 percent level for periods of 12-20 years (which indicates dependence over 24-40-year horizons). In addition, the values of the variance ratio are statistically significantly below unity at the 5 percent level at horizons of 30-34 years. Thus, both test statistics present evidence against the null hypothesis over horizons of 30-34 years. For the second sub-sample, the first-order autocorrelation values are statistically significantly negative at the 5 percent level for periods of 15-18 years (which indicates dependence over 30-36-year horizons), and the values of the variance ratio are statistically significantly below unity at the 5 percent level at horizons of 30-34 years. Thus, our results suggest that accounting for heteroskedasticity in stock returns does not influence the outcomes of the tests based on examining the first-order autocorrelations of $k$-year returns. Regardless of the number of bins in the stratified randomization, the first-order autocorrelation of $k$-year returns remains statistically significantly different from zero at the 5 percent level over periods of 15-17 years for the full sample and the second sub-sample. In contrast, stratification of the sample weakens the evidence against the null hypothesis for the value of the variance ratio. In particular, for the full sample and the stratification with 2, 4, or 5 bins, the variance ratio is not statistically significantly below unity at conventional levels (1 percent and 5 percent). Similarly, for the second sub-sample and the stratification with either 3 or 5 bins, the variance ratio is not significantly below unity at conventional levels. For the first sub-sample, the variance ratio is not significantly below unity regardless of the number of bins used in the stratified randomization.

Consequently, we do not have strong enough evidence to claim that the variance ratio decreases with an increasing investment horizon. Despite that without stratification the variance ratio over horizons of 30-34 years is statistically significantly below unity, stratification of the sample
### TABLE 3
First-order autocorrelations of the k-year log real returns on the Standard and Poor’s Composite Stock Price Index (AC1(k))

| Period, years | 1     | 2     | 3     | 4     | 5     |
|---------------|-------|-------|-------|-------|-------|
| **Panel A: Total sample 1871 to 2011** |       |       |       |       |       |
| 10            | -0.23 (0.14) | -0.15 (0.22) | -0.19 (0.16) | -0.10 (0.29) | -0.06 (0.36) |
| 11            | -0.29 (0.09) | -0.21 (0.15) | -0.25 (0.10) | -0.15 (0.20) | -0.11 (0.26) |
| 12            | -0.38 (0.04) | -0.29 (0.08) | -0.34 (0.05) | -0.23 (0.10) | -0.18 (0.13) |
| 13            | -0.43 (0.03) | -0.34 (0.04) | -0.39 (0.03) | -0.28 (0.06) | -0.23 (0.07) |
| 14            | -0.48 (0.02) | -0.38 (0.03) | -0.44 (0.02) | -0.32 (0.03) | -0.27 (0.04) |
| 15            | -0.50 (0.01) | -0.40 (0.02) | -0.46 (0.01) | -0.34 (0.03) | -0.29 (0.03) |
| 16            | -0.51 (0.01) | -0.40 (0.02) | -0.47 (0.01) | -0.34 (0.02) | -0.28 (0.03) |
| 17            | -0.51 (0.02) | -0.40 (0.02) | -0.46 (0.01) | -0.34 (0.02) | -0.27 (0.02) |
| 18            | -0.50 (0.02) | -0.38 (0.03) | -0.45 (0.02) | -0.33 (0.03) | -0.25 (0.03) |
| 19            | -0.47 (0.03) | -0.34 (0.04) | -0.41 (0.03) | -0.29 (0.04) | -0.21 (0.05) |
| 20            | -0.43 (0.05) | -0.30 (0.07) | -0.37 (0.05) | -0.26 (0.06) | -0.17 (0.09) |
| **Panel B: First sub-sample 1871 to 1956** |       |       |       |       |       |
| 10            | -0.25 (0.17) | -0.11 (0.32) | -0.10 (0.33) | -0.06 (0.39) | -0.26 (0.18) |
| 11            | -0.28 (0.15) | -0.13 (0.29) | -0.12 (0.31) | -0.10 (0.33) | -0.31 (0.14) |
| 12            | -0.36 (0.09) | -0.20 (0.21) | -0.20 (0.22) | -0.19 (0.20) | -0.41 (0.08) |
| 13            | -0.31 (0.13) | -0.13 (0.29) | -0.15 (0.29) | -0.15 (0.26) | -0.37 (0.09) |
| 14            | -0.22 (0.23) | -0.03 (0.46) | -0.06 (0.41) | -0.08 (0.37) | -0.27 (0.12) |
| 15            | -0.13 (0.32) | 0.07 (0.60)  | 0.01 (0.52)  | -0.01 (0.48) | -0.16 (0.20) |
| 16            | -0.03 (0.47) | 0.19 (0.77)  | 0.11 (0.68)  | 0.10 (0.66)  | -0.01 (0.47) |
| 17            | 0.02 (0.52)  | 0.25 (0.83)  | 0.15 (0.74)  | 0.14 (0.72)  | 0.09 (0.74)  |
| 18            | -0.07 (0.41) | 0.19 (0.75)  | 0.06 (0.60)  | 0.05 (0.59)  | 0.07 (0.68)  |
| 19            | -0.16 (0.32) | 0.14 (0.66)  | -0.03 (0.46) | -0.03 (0.46) | 0.05 (0.62)  |
| 20            | -0.17 (0.32) | 0.16 (0.66)  | -0.03 (0.45) | -0.03 (0.46) | 0.11 (0.70)  |
| **Panel C: Second sub-sample 1926 to 2011** |       |       |       |       |       |
| 10            | -0.06 (0.41) | -0.05 (0.41) | 0.09 (0.67)  | -0.03 (0.45) | 0.06 (0.64)  |
| 11            | -0.16 (0.28) | -0.15 (0.27) | -0.00 (0.49) | -0.14 (0.27) | -0.02 (0.45) |
| 12            | -0.23 (0.20) | -0.23 (0.17) | -0.07 (0.35) | -0.22 (0.17) | -0.08 (0.33) |
| 13            | -0.32 (0.11) | -0.33 (0.09) | -0.17 (0.19) | -0.33 (0.08) | -0.16 (0.18) |
| 14            | -0.41 (0.06) | -0.42 (0.05) | -0.27 (0.08) | -0.42 (0.03) | -0.23 (0.07) |
| 15            | -0.46 (0.04) | -0.47 (0.03) | -0.34 (0.05) | -0.48 (0.02) | -0.29 (0.03) |
| 16            | -0.53 (0.03) | -0.54 (0.01) | -0.42 (0.03) | -0.54 (0.01) | -0.37 (0.01) |
| 17            | -0.53 (0.04) | -0.54 (0.02) | -0.45 (0.03) | -0.55 (0.01) | -0.41 (0.01) |
| 18            | -0.50 (0.06) | -0.51 (0.03) | -0.43 (0.04) | -0.51 (0.01) | -0.42 (0.01) |
| 19            | -0.45 (0.09) | -0.46 (0.05) | -0.40 (0.06) | -0.46 (0.01) | -0.43 (0.01) |
| 20            | -0.40 (0.13) | -0.43 (0.08) | -0.36 (0.09) | -0.41 (0.03) | -0.42 (0.02) |

Note: These estimates are obtained using the randomization method with stratification (when the number of bins is greater than one). The estimates are corrected for bias under the null hypothesis. The values in the brackets report the p-values of a one-tailed test of the hypothesis $H_0: \text{AC}1(k) = 0$. Bold text indicates values that are statistically significant at the 5% level.
**TABLE 4**  
Variance ratios of the k-year log real returns on the Standard and Poor’s Composite Stock Price Index (VR(k))

| Number of bins | 1    | 2    | 3    | 4    | 5     |
|----------------|------|------|------|------|-------|
| **Panel A: Total sample 1871 to 2011** |
| 20  | 0.75 (0.21) | 0.87 (0.32) | 0.80 (0.22) | 0.93 (0.39) | 1.00 (0.50) |
| 22  | 0.71 (0.17) | 0.84 (0.28) | 0.77 (0.17) | 0.90 (0.34) | 0.97 (0.45) |
| 24  | 0.66 (0.12) | 0.79 (0.21) | 0.71 (0.11) | 0.86 (0.25) | 0.93 (0.35) |
| 26  | 0.62 (0.09) | 0.76 (0.16) | 0.67 (0.07) | 0.83 (0.18) | 0.90 (0.26) |
| 28  | 0.58 (0.06) | 0.73 (0.12) | 0.64 (0.04) | 0.80 (0.12) | 0.87 (0.19) |
| 30  | 0.58 (0.05) | 0.73 (0.11) | 0.63 (0.03) | 0.80 (0.10) | 0.87 (0.16) |
| 32  | 0.59 (0.05) | 0.75 (0.11) | 0.64 (0.03) | 0.81 (0.11) | 0.88 (0.17) |
| 34  | 0.60 (0.05) | 0.76 (0.11) | 0.65 (0.03) | 0.82 (0.11) | 0.89 (0.17) |
| 36  | 0.62 (0.06) | 0.78 (0.13) | 0.67 (0.03) | 0.83 (0.13) | 0.91 (0.21) |
| 38  | 0.65 (0.08) | 0.81 (0.16) | 0.70 (0.04) | 0.86 (0.18) | 0.93 (0.28) |
| 40  | 0.67 (0.09) | 0.84 (0.18) | 0.72 (0.05) | 0.87 (0.20) | 0.95 (0.32) |
| **Panel B: First sub-sample 1871 to 1956** |
| 20  | 0.66 (0.11) | 0.82 (0.21) | 0.84 (0.18) | 0.91 (0.30) | 0.71 (0.07) |
| 22  | 0.67 (0.10) | 0.82 (0.20) | 0.84 (0.16) | 0.90 (0.27) | 0.69 (0.05) |
| 24  | 0.65 (0.06) | 0.81 (0.13) | 0.82 (0.09) | 0.87 (0.17) | 0.66 (0.02) |
| 26  | 0.70 (0.09) | 0.85 (0.18) | 0.86 (0.14) | 0.90 (0.25) | 0.69 (0.03) |
| 28  | 0.74 (0.13) | 0.89 (0.25) | 0.90 (0.23) | 0.94 (0.33) | 0.73 (0.04) |
| 30  | 0.79 (0.18) | 0.94 (0.34) | 0.93 (0.31) | 0.97 (0.40) | 0.77 (0.06) |
| 32  | 0.84 (0.26) | 0.99 (0.46) | 0.97 (0.43) | 1.01 (0.53) | 0.84 (0.12) |
| 34  | 0.86 (0.26) | 0.99 (0.48) | 0.98 (0.43) | 1.01 (0.53) | 0.87 (0.13) |
| 36  | 0.85 (0.21) | 0.97 (0.40) | 0.95 (0.34) | 0.98 (0.42) | 0.87 (0.08) |
| 38  | 0.84 (0.16) | 0.96 (0.33) | 0.94 (0.27) | 0.97 (0.33) | 0.89 (0.08) |
| 40  | 0.85 (0.14) | 0.96 (0.31) | 0.94 (0.24) | 0.97 (0.30) | 0.92 (0.11) |
| **Panel C: Second sub-sample 1926 to 2011** |
| 20  | 0.98 (0.48) | 1.01 (0.51) | 1.20 (0.79) | 1.10 (0.67) | 1.18 (0.81) |
| 22  | 0.94 (0.44) | 0.97 (0.46) | 1.17 (0.76) | 1.05 (0.59) | 1.15 (0.77) |
| 24  | 0.90 (0.38) | 0.91 (0.38) | 1.12 (0.71) | 0.98 (0.47) | 1.10 (0.69) |
| 26  | 0.83 (0.28) | 0.84 (0.26) | 1.04 (0.59) | 0.89 (0.30) | 1.03 (0.56) |
| 28  | 0.78 (0.19) | 0.78 (0.15) | 0.98 (0.44) | 0.82 (0.18) | 0.97 (0.42) |
| 30  | 0.76 (0.14) | 0.77 (0.10) | 0.94 (0.32) | 0.79 (0.13) | 0.95 (0.34) |
| 32  | 0.74 (0.07) | 0.74 (0.04) | 0.90 (0.17) | 0.76 (0.07) | 0.92 (0.19) |
| 34  | 0.74 (0.04) | 0.74 (0.02) | 0.88 (0.09) | 0.76 (0.04) | 0.91 (0.11) |
| 36  | 0.76 (0.04) | 0.76 (0.01) | 0.89 (0.09) | 0.78 (0.03) | 0.92 (0.12) |
| 38  | 0.79 (0.07) | 0.79 (0.02) | 0.91 (0.14) | 0.81 (0.04) | 0.93 (0.16) |
| 40  | 0.82 (0.07) | 0.81 (0.02) | 0.92 (0.15) | 0.84 (0.03) | 0.93 (0.17) |

Note: These estimates are obtained using the randomization method with stratification (when the number of bins is greater than one). The estimates are corrected for bias under the null hypothesis. The values in the brackets report the p-values of a one-tailed test for the hypothesis $H_0 : VR(k) = 1$. Bold text indicates values that are statistically significant at the 5% level.
suggests that this effect can be attributed to the historical pattern of heteroskedasticity (that is, the existence of periods of high and low variance). Thus, our results cannot fully support the anecdotal evidence that states that the stock market is safer for long-term investors. Nevertheless, we do have strong enough evidence to reject the random walk hypothesis in stock prices over periods of approximately 15-17 years. This evidence is based on the first-order autocorrelation of multi-year returns. However, our results suggest that the departure from the random walk over very long horizons is primarily a phenomenon of the post-1926 period.

III.3 Robustness Tests

To check the robustness of our findings regarding the statistical significance of secular mean reversion, we conducted a series of robustness checks, the full results of which are not reported in this paper to save space. These additional robustness tests are described below.

First, the results reported in this section are obtained using the annual data provided by Robert Shiller. Specifically, these data are annual series of (average) January values of the real Standard and Poor’s Composite Stock Price Index. Hence, the results obtained in this section might be affected by seasonality. To test the seasonality problem, we used the monthly data instead and obtained virtually the same levels of statistical significance of the mean-reverting behavior over very long horizons.

Second, Robert Shiller uses the CPI to adjust the nominal returns for inflation. We tested whether our evidence of mean reversion depends on the choice of deflator used to construct real stock returns. For this purpose, we constructed the real stock returns using the GDP deflator and the value of the consumer bundle. We found that regardless of the choice of a deflator, the evidence on mean reversion remains intact.

Third, since Kim et al. (1991) demonstrated that the mean-reversion in the study by Fama and French (1988b) is primarily a phenomenon of the pre-World War II period, which is included in both of our sub-samples, we tested whether there is evidence of mean-reversion in the post-1940 period. We found that the evidence is weaker (which is reasonable since the sample size decreases) but is still statistically significant at the 10 percent level.

Fourth, instead of the first-order autocorrelation of multi-year returns test statistic, suggested by Fama and French (1988b), we used the test statistic suggested by Jegadeesh (1991). In particular, instead of regression (3), we used the following regression

\[ r_t = a(k) + b(k) r_{t-k} + \epsilon_{t, t+k} . \]  

(4)

Note that in this regression, the stock market return in year \( t \) is predicted using the aggregate return over the preceding \( k \) years. Using this regression, we were also able to reject the random walk hypothesis in stock prices over very long horizons in the post-1926 period.

Finally, instead of using the data provided by Robert Shiller, we used the real annual returns on large-cap stocks provided by Kenneth French over the period from 1927 to 2012. Again, we found that the values of the first-order autocorrelation of multi-year returns are statistically significantly negative over periods of 15-18 years.

---

6 We thank Ole Gjølberg for bringing our attention to this.
7 We thank an anonymous referee for suggesting this.
8 The data on the GDP deflator and the consumer bundle are downloaded from www.measuringworth.com. The value of the consumer bundle is defined as the average annual expenditures of consumer units.
9 We thank an anonymous referee for suggesting this.
10 See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. We used the large-cap stocks because the S&P Composite is a large-cap index.
Thus, on the basis of the results from numerous robustness tests, we conclude that our evidence on secular mean reversion is robust to the choice of data, deflator, sample period, and test statistics.

IV. TESTING THE LONG-HORIZON RETURN PREDICTABILITY

IV.1 Motivation

The results of the tests performed in the preceding section allow us to reject the hypothesis that the real S&P Composite Stock Price Index follows a random walk. Rather surprisingly, considering a seemingly insufficient span of available historical observations of the returns on the stock index, convincing evidence against the random walk is present over long-lasting periods of approximately 15-17 years. That is, our tests support the alternative hypothesis that there is serial dependence in stock returns. This leads us to the following question: what type of serial dependence is present? In other words, what is the alternative hypothesis? Typically, a statistically significant decrease in the variance ratio with an increasing investment horizon (this effect is sometimes termed ‘variance compression’) is interpreted as evidence of mean reversion. Unfortunately, the evidence of mean reversion based on the variance ratio test does not appear to be sufficiently strong under a stratified randomization of the data. However, variance compression seems to be a sufficient, but likely not necessary, condition for mean reversion. Fortunately, in addition to the variance ratio, we have another test statistic, namely, the first-order autocorrelation of multi-year returns. The significance of this test statistic is unaffected by the choice of randomization method. The presence of values of the autocorrelation of $k$-year returns that are statistically significantly below zero suggests mean-reverting behavior in stock prices. Specifically, a given change in price over first $k$ years tends to be reversed over the next $k$ years by a predictable change in the opposite direction. For the full sample period, evidence for mean reversion comes from the negative and statistically significant values of the first-order autocorrelations, particularly for periods of 15, 16, and 17 years.

Considering the above, the results reported in the previous section suggest the presence of long-term mean reversion over periods of approximately 15-17 years in the real Standard and Poor’s Composite Stock Price Index. In this case, if the pattern of first-order autocorrelation of multi-year returns suggests the presence of mean reversion over a horizon of $2k$ years, there should be some degree of predictability of multi-year returns over half of this horizon, that is, over a period of $k$ years. Indeed, regression (3) is a predictive regression. To demonstrate the predictability of multi-year returns, Figure 2 presents a scatter plot of $r_{t,t+15}$ against $r_{t-15,t}$, for the returns on the real Standard and Poor’s Composite Stock Price Index for the full sample period from 1871 to 2011. In addition, a regression line is fitted through these data points. The scatterplot clearly suggests a tendency for the past 15-year returns to predict future 15-year returns. The regression line has a strongly negative slope, and the $R^2$ statistic is 42 percent.

However, if we use the full sample period to estimate the first-order autocorrelation of multi-year returns, our estimate measures the degree of in-sample (IS) predictability. The issue here is the well-known possibility that in-sample predictability might be spurious (for example, it could appear as a result of data mining) and not hold out-of-sample (OOS) (see, for example, Bossaerts and Hillion, 1999, Goyal and Welch, 2003, and Welch and Goyal, 2008). To guard against data mining, in this section, we assess the performance of the OOS forecast based on the mean-revering model given by regression (3). In addition to the mean-reverting model, we use several other competing predictive models. We demonstrate that in the OOS tests, the mean-reverting model and a few other predictive models perform statistically significantly better.
than the naive historical-mean model. In addition, we demonstrate that the advantages of the predictive models translate into significant utility gains.

IV.2 Methodology for Assessing the Performance of OOS Forecasts

Our OOS recursive forecasting procedure is as follows. The initial IS period \([1, m]\), \(m < n\), is used to estimate regression (3) for different period lengths \(k \in [10, 20]\) years. In this manner, we estimate a number of autocorrelations of \(k\)-year returns, \(AC1(k)\). Then, we perform the bias adjustment of \(AC1(k)\). Next, we select the value of \(k = k_1\) that produces the lowest estimate of the bias-adjusted autocorrelation. That is,

\[
k_1 = \arg \min_{k \in [10, 20]} AC1(k).
\]

Presumably, over the initial IS period, the evidence of mean reversion is strongest over the period of \(k_1\) years. Subsequently, the estimated coefficients from regression (3) with \(k_1\) are used to compute the first \(k_1\)-year-ahead return forecast for the period \([m + 1, m + k_1]\). We then expand our IS period by one year (it becomes \([1, m + 1]\)), perform the selection of the \(k_2\) at which the evidence of mean reversion is strongest over the second IS period, and compute the OOS forecast for the period \([m + 2, m + k_2 + 1]\). We repeat the procedure, each time increasing our IS window by one year, until we compute the last \(k_l\)-year-ahead return for the period \([n - k_l + 1, n]\).

It is worth emphasizing that our OOS forecasting procedure is free from look-ahead bias, as in forecasting the return for period \([m + j, m + k_j + j - 1]\), \(j \geq 1\), we use only information that is available at time \(m + j - 1\). Note that since we are conducting a long-horizon forecast, in performing the recursive forecasting procedure, we need not only update the estimates of

Fig. 2. This figure shows a scatter plot of \(r_{t+15}\) against \(r_{t-15}\) for the log real Standard and Poor’s Composite Stock Price Index for the period from 1871 to 2011. In addition, a regression line is fit through these data points. The goodness of fit, as measured by \(R^2\), amounts to 42%.
the coefficients of regression (3), but we also first need to update the optimal length of the prediction period, $k$. Observe that, to avoid look-ahead bias, the optimal length of the prediction period, $k$, is also determined using only information that is available at the end of each IS period. Thus, our OOS recursive forecasting procedure updates all values of the model parameters and is able to adapt to changing conditions in the time series. For example, it can accommodate the possibility that the period of mean reversion is monotonically changing over time.\textsuperscript{11}

A common approach to assess the performance of an OOS forecast in the empirical literature is to run a ‘horse-race’ among several competing predictive models. A standard criterion by which to compare two alternative predictive models is to compare their mean squared prediction errors (MSPE). A comparison of the mean squared prediction errors of two alternative models has a long tradition in evaluating which of two models has better forecasting ability; see McCracken (2007) and references therein. In our study, we run OOS horse races involving the mean-reverting model (MR), the historical-mean model (HM), Robert Shiller’s model (PE10), which uses the cyclically adjusted price-to-earnings ratio as a predictor of long-horizon returns, the model that uses the price-to-dividends ratio (PD) as a predictor, and the model that uses the long-term bond yield (LTY) as a predictor. These models are given by

$$
\text{MR} : r_{t+k} = a(k) + b(k) r_{t-k, t} + \varepsilon_{t, t+k},
$$

$$
\text{PE10} : r_{t+k} = a(k) + b(k) p {e10}, + \varepsilon_{t, t+k},
$$

$$
\text{PD} : r_{t+k} = a(k) + b(k) pd, + \varepsilon_{t, t+k},
$$

$$
\text{LTY} : r_{t+k} = a(k) + b(k) lty, + \varepsilon_{t, t+k},
$$

$$
\text{HM} : r_{t+k} = a(k) + \varepsilon_{t, t+k},
$$

where $p {e10}$ is the natural log of the ratio of the price to the 10-year moving average of earnings (this ratio is usually denoted CAPE or PE10), $pd$ is the natural log of the price-to-dividends ratio, and $lty$ is the natural log of the long-term bond yield. The data for the price-to-earnings ratio, price-to-dividends ratio, and the long-term bond yield are also provided by Robert Shiller.

Robert Shiller’s model was introduced by Campbell and Shiller (1998) and further popularized and developed by Shiller (2000) and Campbell and Shiller (2001). Shiller’s model is based on a simple mean reversion theory that states that when stock prices are very high relative to recent earnings, then prices will eventually fall in the future to return the price-to-earnings ratio to a more normal historical level. Using this model, Campbell and Shiller (1998) predicted the stock market crash of 2000 on the basis of an unreasonably high PE10 ratio. Since that time, Shiller’s model has been extremely popular among practitioners. Originally, Campbell and Shiller (1998), Shiller (2000), and Campbell and Shiller (2001) used this model to forecast future 10-year returns. Yet, Asness (2003) demonstrated that the PE10 ratio is a good predictor of the future returns over periods from 10 to 20 years.\textsuperscript{12} Thus, Shiller’s model represents a natural competitor to our long-term mean-reverting model.

\textsuperscript{11}Recall that the results presented in the previous section indicate that the period of the long-term mean reversion seems to have been increasing over time. In particular, during the first sub-sample, the evidence of mean reversion is strongest over horizons of approximately 24-26 years (judging by the values of the most statistically significant first-order autocorrelation and variance ratio). In contrast, during the second sub-sample, the evidence of mean reversion is strongest over horizons of approximately 34-36 years. This results in the observation that over the full sample period, the evidence of mean reversion is strongest over horizons of approximately 30-34 years.

\textsuperscript{12}This conclusion is made on the basis of studying the $R^2$ of the predictive regression at different forecasting horizons. It should be noted, however, that in estimating the coefficient in front of the predictor
The model that uses the price-to-dividends ratio as a predictor of future returns was presented by Fama and French (1988a). This model is also based on a simple mean reversion theory that states that if the price-to-dividends ratio is unusually high or low, then this ratio tends to return to its long-run historical mean. The motivation for the model that uses the long-term bond yield as a predictor is based on the simple idea that stocks and long-term bonds are two major competing assets. Therefore, simple logic suggests that the changes in the long-term bond yield must be highly correlated with the changes in the stock market earnings yield (earnings-to-price ratio). If, for example, the bond yield increases, stock prices should decrease and the stock market earnings yield increases. The so-called ‘Fed model’ postulates that the stock market’s earnings yield should be approximately equal to the long-term bond yield. Empirical support for this model is found in studies by Lander, Orphanides, and Douvogiannis (1997), Koivu, Pennanen, and Ziemb (2005), Berge et al. (2008), and Maio (2013).

The historical-mean model can be interpreted as a reduced version of any other predictive model. This model uses the historical average of \(k\)-year returns to predict the return over the next \(k\) years. It is worth emphasizing that Welch and Goyal (2008) also employed predictive models that use the price-to-earnings ratio, price-to-dividends ratio, and the long-term bond yield. They found in OOS tests that these models perform worse than the historical-mean model. However, these authors used an increasing forecast horizon of only up to 5 years. In our study, the goal is to compare the OOS forecast accuracy from these models over horizons longer than 10 years.

Now, we turn to the formal presentation of our test statistic that is employed to assess the performance of OOS forecasts provided by two competing models. Let \(r_{AC,t+k}\), \(t > m\), be the actual \(k\)-year returns and \(r_{mod1,t+k}\) and \(r_{mod2,t+k}\) be the OOS forecast of the \(k\)-year returns provided by models 1 and 2. To compute the test statistic, we first compute the OOS prediction errors of the two competing models

\[
\epsilon_{mod1,t+k} = r_{mod1,t+k} - r_{AC,t+k}, \quad \epsilon_{mod2,t+k} = r_{mod2,t+k} - r_{AC,t+k}.
\]

Our test statistic is the ratio of the MSPE of model 1 to the MSPE of model 2

\[
\text{MSPE-R} = \frac{1}{T-m} \sum_{t=m+1}^{T} \left( \frac{\epsilon_{mod1,t+k}}{\epsilon_{mod2,t+k}} \right)^2,
\]

where \(T - m\) is the number of OOS forecasted \(k\)-year returns.\(^{13}\) The null hypothesis in this test is that the forecast provided by model 2 is not better than the forecast provided by model 1. Specifically, under the null hypothesis, the MSPE of model 1 is less than or equal to the MSPE of model 2. Formally, \(H_0 : \text{MSPE-R} \leq 1\). Consequently, we reject the null hypothesis when the actual estimate of the MSPE ratio is significantly above unity. In our tests, model 1 is always the historical-mean model. Therefore, the outcome of our tests is whether a predictive model can ‘beat’ the historical-mean model (a similar approach was used by Goyal and Welch (2003), Welch and Goyal (2008), and many others).

If two alternative prediction errors are assumed to be Gaussian, serially uncorrelated, and contemporaneously uncorrelated, then an MSPE-R statistic under the null hypothesis has the usual \(F\)-distribution.\(^{14}\) However, in our case, the assumptions listed above are not met. First, because we use overlapping multi-year returns, the prediction errors of all our models are serially correlated. Second, since the historical-mean model is the reduced version of any other and its significance level, Asness (2003) did not account for the estimation biases discovered by Cavanagh et al. (1995) and Stambaugh (1999).

\(^{13}\)Note that \(k\) is not constant but a variable that is exogenously determined by our recursive forecasting procedure. We suppress its dependence on time to simplify the notation.

\(^{14}\)In this case, testing the null hypothesis largely corresponds to the standard \(F\)-test of equal forecast error variances.

© 2016 Board of Trustees of the Bulletin of Economic Research and John Wiley & Sons Ltd
predictive model, the prediction errors of the historical-mean models and any other predictive model are contemporaneously correlated. Finally, the assumption of Gaussian errors also seems to be inappropriate. One possibility for obtaining correct statistical inference in this case is to perform asymptotically valid tests in the spirit of the seminal tests by Diebold and Mariano (1995). However, because we use relatively small samples, and because of the variable length $k$ of the prediction horizon in our forecasting procedure, to compute the p-value of the MSPE ratio, we employ a bootstrap method.

Our bootstrap method closely follows Welch and Goyal (2008). In this method, we assume that the returns are serially independent, whereas the log of the PE10, the log of the PD, and the log of LTY follow a first-order autoregressive (AR(1)) process. Therefore, the data-generating process is assumed to be

$$
\begin{align*}
r_t &= \mu + u_t, \\
pe10_t &= \alpha_1 + \beta_1 pe10_{t-1} + w_t, \\
pd_t &= \alpha_2 + \beta_2 pd_{t-1} + z_t, \\
lty_t &= \alpha_3 + \beta_3 lty_{t-1} + e_t.
\end{align*}
$$

(10)

In this case, the return series, $r_t$, follows a random walk and a bootstrapped resample is generated using a nonparametric bootstrap method. In particular, a random resample $(r_1^*, r_2^*, \ldots, r_n^*)$ is generated by drawing with replacement from the original series $(r_1, r_2, \ldots, r_n)$. In contrast, a bootstrapped resample of any other predictive variable is generated using a semi-parametric bootstrap method. The construction of a bootstrapped resample for the log of the PE10 series, $pe10_t$, is performed as follows. First, the parameters $\alpha_1$ and $\beta_1$ are estimated by OLS using the full sample of observations, with the residuals stored for resampling. Next, to generate a random resample $(pe10_{t-1}^*, pe10_{t-2}^*, \ldots, pe10_t^*)$ we select an initial observation $pe10_t^*$ from the actual data at random. Then, a series is generated using the AR(1) model and by drawing $w_t^*$ with replacement from the residuals. The construction of a bootstrapped resample for the log of the PD and the LTY series is done in a similar manner.

Now, we describe how we compute the MSPE-R statistic and its p-value. First, using the original series $(r_1, r_2, \ldots, r_n)$, we employ the recursive forecasting procedure described above to obtain the OOS forecasts of the mean-reverting model. Note that one of the outcomes of our recursive forecasting procedure is a sequence of lengths of prediction periods $(k_1, k_2, \ldots, k_l)$. Second, using the same sequence of lengths of prediction periods, we obtain the OOS forecasts of all other models. Thereafter, we compute the mean squared prediction errors and, then, the MSPE-R statistic. We then bootstrap the original series to obtain random resamples. The next crucial step is to generate a sequence of lengths of prediction periods $(k_1^*, k_2^*, \ldots, k_l^*)$. All of this is repeated 10,000 times, each time running the recursive forecasting procedures and obtaining an estimate of MSPE-R$^*$. In this manner, we estimate the sampling distribution of the MSPE-R statistic under the null hypothesis. Finally, to estimate the significance level, we count how many times the computed value of the MSPE-R$^*$ after bootstrapping is above the value of

---

15 Note that in this case, the historical-mean model is a version of the random walk hypothesis.

16 It should be noted, however, that our data-generating process assumes no contemporaneous correlation between the stock return and a predictive variable. In the actual data, there is a small but statistically significant correlation between the returns and the price-to-earnings (as well as the price-to-dividends) ratio. To check the robustness of our findings, we also implemented another bootstrap method that retains the historical correlations between the data series. We found that both bootstrap methods deliver similar p-values for our test statistic.

17 Note that here the recursive forecasting procedures for all the models use the exogenously determined sequence of lengths of prediction periods.
the actual estimate of the MSPE-R. In other words, under the null hypothesis, we compute the probability of obtaining a more extreme value of the MSPE ratio than the actual estimate.\textsuperscript{18}

It is not clear what method should be used to generate a sequence of lengths of prediction periods for each bootstrap simulation. To the best of our knowledge, there are no similar forecasting procedures in the relevant scientific literature. Therefore, we entertain four different methods listed below. In the first method, we always use the original sequence of lengths of prediction periods ($k_1, k_2, \ldots, k_l$). In the second and third methods, a generated sequence ($k^*_1, k^*_2, \ldots, k^*_l$) is a bootstrapped version of the original sequence. Whereas in the second method, we use the nonparametric bootstrap, in the third method, we use the semi-parametric bootstrap. In the semi-parametric bootstrap, we assume that the length of a prediction period is a linear function of time.\textsuperscript{19} In the fourth method, a sequence of lengths of prediction periods is endogenously determined by the recursive forecasting procedure on the basis of the bootstrapped series ($r^*_1, r^*_2, \ldots, r^*_n$). We find that the first three methods produce virtually identical p-values, whereas the fourth method produces notably lower p-values. When we report the p-values of the MSPE-R statistic, we use the highest p-values. Thus, our statistical inference is based on the ‘worst-case scenario’ for the rejection of the null hypothesis. In other words, if we can reject the null in the ‘worst-case scenario’, we would reject it for any other case.

\textbf{IV.3 Empirical Results on the Performance of OOS Forecasts}

Our OOS forecast begins 50 years after the data are available, that is, in 1921, and ends in 1997 with the last forecast for the 15-year period from 1997 to 2011. To check the robustness of findings, we divide the total OOS period into two equal OOS sub-periods, the first from 1921 to 1959 and the second from 1959 to 1997. As in Goyal and Welch (2003), we employ a simple graphical diagnostic tool that makes it easy to understand the relative performance of two competing forecasting models. In particular, to monitor the predictive power of the unrestricted model relative to the that of the restricted model, Goyal and Welch (2003) suggested using the cumulative difference between the MSPE of the restricted model (the HM model in our case) and the MSPE of the unrestricted model

$$CUDIF_t = \sum_{i=m}^{t} (\varepsilon_{mod_1}^{i+k})^2 - (\varepsilon_{mod_2}^{i+k})^2.$$ 

From a visual examination of the graph of $CUDIF_t$, it is easy to understand in which periods the unrestricted model performs better than the restricted model. Specifically, in periods when the cumulative MSPE difference increases, the unrestricted model provides better predictions; in periods when it decreases, the unrestricted model has worse predictive performance than the restricted model.

Figure 3 shows the performance of the unrestricted models versus the performance of the restricted (historical-mean) model. Specifically, the left panels in the figure plot the actual $k$-year-ahead returns versus the OOS forecasted $k$-year-ahead returns produced by the unrestricted and restricted models. The right panels in the figure plot the cumulative difference between

\textsuperscript{18}Note again that we compute p-values of a one-tailed test in this manner.

\textsuperscript{19}Indeed, for our OOS period from 1921 to 1997, the length of a prediction period is almost monotonically stepwise increasing from 10 to 15 years. The goodness of fit to the linear function, as measured by $R^2$, amounts to 73 percent. To perform the semi-parametric bootstrap, first, we estimate the simple linear trend model for the original sequence of lengths of prediction periods ($k_1, k_2, \ldots, k_l$), with the residuals stored for resampling. Thereafter, to generate a random resample of the sequence of lengths of prediction periods, we select the original initial prediction period, $k_1$. The rest of the sequence is generated using the estimated linear trend model by drawing the error terms from the residuals with replacement.
Fig. 3. Performance of the unrestricted models versus the performance of the restricted (historical-mean) model. The left panels plot the actual $k$-year-ahead returns (solid black line) versus the $k$-year-ahead returns forecasted OOS by the unrestricted (thick solid grey line) and restricted (dotted black line) models. The initial IS period is from 1871 to 1920, which covers a span of 50 years. The OOS forecast begins in 1921 and ends in 1997 with the last forecast for the 15-year period from 1997 to 2011. The right panels plot the cumulative difference between the MSPE of the restricted model and the MSPE of the unrestricted model.
TABLE 5

The values of the MSPE-R statistic with corresponding p-values (in brackets)

| OOS period | HM to MR | HM to PE10 | HM to PD | HM to LTY |
|------------|----------|------------|----------|-----------|
| 1921-1997 | 1.35 (0.01) | 1.44 (0.01) | 1.33 (0.03) | 0.44 (0.98) |
| 1921-1959 | 0.86 (0.38) | 1.14 (0.10) | 1.38 (0.04) | 0.37 (0.97) |
| 1959-1997 | 2.14 (0.01) | 1.78 (0.01) | 1.28 (0.09) | 0.50 (0.92) |

Note: An MSPE-R statistic is a ratio of the mean squared prediction error of the restricted (historical-mean) model to the mean squared prediction error of an unrestricted model. The four competing unrestricted models are as follows: the mean-reverting (MR) model, the price to 10-year moving average of earnings (PE10) model, the price-dividends (PD) model, and the long-term bond yield model (LTY). For example, column HM to MR reports the values of the ratio of the MSPE of the historical-mean model to the MSPE of the mean-reverting model. The estimated p-values of the MSPE ratios are based on performing 10,000 bootstraps. Bold text indicates values that are statistically significantly above unity at the 5 percent level.

The p-values of the MSPE-R statistic demonstrate that over the full OOS period, 3 out of 4 unrestricted models performed statistically significantly better (at the 5 percent level) than the restricted model. These unrestricted models are as follows: the mean-reverting model, the price-earnings model, and the price-dividends model. However, over the first OOS sub-period, only the price-dividends model performed statistically significantly better than the historical-mean model. In contrast, over the second OOS sub-period, only the mean-reverting and price-earnings models showed evidence of superior forecasting accuracy compared to that of the historical-mean model. Our results advocate that the model using the long-term bond yield as predictor performed substantially worse than all of the other competing models. Our results on the predictive ability of the long-term bond yield support the conclusions reached in the studies by Estrada (2006) and Estrada (2009). Specifically, Estrada argued that the predictive ability of the long-term bond yield is supported by data in the post-1960 period only. Prior to 1960, there is no empirical support for a model that uses the long-term bond yield as a predictor of stock returns.

The graphs of the cumulative difference between the MSPE of the restricted (historical-mean) model and the unrestricted model allow us to determine in which historical periods one model performed better than the other. Visual inspection of these graphs reveals the following observations. The price-dividends model performed relatively well only until circa 1970. Thereafter, the accuracy of the forecast provided by the price-dividends model was substantially worse than that of the historical-mean model. Both the mean-reverting and price-earnings models performed significantly better than the historical-mean model over the period 1960-1990. From circa 1990, the price-earnings model lost its advantage over the historical-mean model. Beginning from circa 1980, the mean-revering model performed substantially better than all other competing models. Only over the decade of the 1950s did the mean-reverting model perform notably worse than the historical-mean model.

IV.4 Economic Significance of Return Predictability

In the preceding subsection, we found statistically significant evidence of long-term predictability of stock returns. This evidence was obtained by comparing the MSPE of the predictive model with the MSPE of the historical-mean model. However, over the total OOS period, the ratios of

20In all empirical studies that demonstrate the predictive ability of the long-term bond yield, the sample period begins after 1960. In this case, if, for example, the initial IS period is 1960-1980, then over the OOS period 1980-2010, one finds evidence of OOS predictability of stock returns using the long-term bond yield.
the MSPE of the restricted model to the MSPE of the unrestricted model are not substantially above unity. This raises the important question of whether they are economically meaningful. In other words, statistical significance is not equivalent to economic significance.

To estimate the economic significance of return predictability, we closely follow the methodology employed in the studies by Fleming et al. (2001), Campbell and Thompson (2008), and Kirby and Ostdiek (2012). We consider an investor who, at time $t$, allocates a proportion $y_t$ of his wealth to the stock market index and a proportion $(1 - y_t)$ to the risk-free asset. The investor revises the composition of his portfolio at time $t + q$; that is, after $q$ years, $q \geq 1$. The investor’s return over period $(t, t + q)$ is given by

$$ R_{t,t+q} = y_t r_{t,t+q} + (1 - y_t) r_{t,t+q}^\text{free}, $$

where $r_{t,t+q}$ and $r_{t,t+q}^\text{free}$ are the stock market return and the risk-free rate of return over period $(t, t + q)$.

We assume that the investor behaves according to a mean-variance utility function, which can be considered a second-order approximation of the investor’s true utility function. As a result, the investor’s realized utility over period $(t, t + q)$ can be written as

$$ u(R_{t,t+q}) = y_t (r_{t,t+q} - r_{t,t+q}^\text{free}) - \frac{1}{2} y^2 \sigma_{t,t+q}^2, $$

where $\sigma_{t,t+q}$ is the volatility of the stock market index over period $(t, t + q)$ and $\gamma$ is the investor’s coefficient of risk aversion. The investor’s total realized utility is found as the sum of single-period utilities

$$ U(R) = \sum_{i=1}^{n} u(R_{t,i+q}), \quad t = (i - 1) \times q, $$

where $n = \frac{T}{q}$ is the number of periods of length $q$ from time 0 to time $T$ (the end of the investment horizon).

The investor’s optimal proportion $y_t$, which maximizes the expected utility, is given by (see Bodie et al., 2007, Chapter 7)

$$ y_t = \frac{1}{\gamma} \left( \frac{E[r_{t,t+q}] - r_{t,t+q}^\text{free}}{\sigma_{t,t+q}^2} \right), $$

where $E[r_{t,t+q}]$ and $\sigma_{t,t+q}$ are the expected return and volatility over $(t, t + q)$ that need to be forecasted at time $t$. Expected returns are forecasted using two competing models, 1 and 2. Specifically, $\hat{r}_{t,t+q}^{\text{mod}_1}$ and $\hat{r}_{t,t+q}^{\text{mod}_2}$ denote the return forecasts provided by models 1 and 2, respectively. Since we do not have a specific predictive model to forecast the volatility, the volatility over $(t, t + q)$ is forecasted using the historical-mean model for volatility. Formally,

$$ \hat{\gamma}_t^{\text{mod}_1} = \frac{1}{\gamma} \left( \frac{\hat{r}_{t,t+q}^{\text{mod}_1} - r_{t,t+q}^\text{free}}{\hat{\sigma}_{t,t+q}^2} \right), \quad \hat{\gamma}_t^{\text{mod}_2} = \frac{1}{\gamma} \left( \frac{\hat{r}_{t,t+q}^{\text{mod}_2} - r_{t,t+q}^\text{free}}{\hat{\sigma}_{t,t+q}^2} \right), $$

where $\hat{\sigma}_{t,t+q}$ denotes the forecasted volatility.

It is important to observe that our predictive models forecast the stock market returns for a period of $k \geq 10$ years. Because, generally, $q \neq k$ (most often $q < k$), the $q$-year forecasted returns for model $i \in \{1, 2\}$ are computed as

$$ \hat{r}_{t,t+k}^{\text{mod}_i} = \hat{r}_{t,t+k}^{\text{mod}_i} \times \frac{q}{k}, $$

where $\hat{r}_{t,t+k}^{\text{mod}_i}$ is the $k$-year return forecast provided by model $i$. © 2016 Board of Trustees of the Bulletin of Economic Research and John Wiley & Sons Ltd
The table reports the performance of alternative predictive models and the annual fees the investor is willing to pay to switch from the historical-mean model to another predictive model.

### Table 6

| Forecasting model       | Sharpe ratio | Basis point fees |
|-------------------------|--------------|------------------|
| **Panel A: Portfolio rebalancing once a year** |              |                  |
| Historical-Mean         | 0.35         | 0                |
| Mean-Reverting           | 0.42         | 46               |
| Price-Earnings           | 0.45         | 77               |
| Price-Dividends          | 0.39         | 30               |
| Bond Yield               | 0.32         | -20              |
| **Panel B: Portfolio rebalancing once in 15 years** |              |                  |
| Historical-Mean         | 0.35         | 0                |
| Mean-Reverting           | 0.40         | 47               |
| Price-Earnings           | 0.35         | 1                |
| Price-Dividends          | 0.39         | 37               |
| Bond Yield               | 0.21         | -129             |

Note: The performance is measured using the Sharpe ratio. The annual fees are measured in basis points.

As before, model 1 in our study is the historical-mean model. The economic significance of return predictability is measured by equating the two total realized utilities associated with two alternative forecasting models

\[
\sum_{t=1}^{n} u \left( R_{t,t+q}^{\text{mod}_1} \right) = \sum_{t=1}^{n} u \left( R_{t,t+q}^{\text{mod}_2} - q \times \Delta \right),
\]

where \( \Delta \) denotes the annual fees the investor is willing to pay to switch from predictive model 1 to predictive model 2. Whereas Fleming et al. (2001) and Kirby and Ostdiek (2012) used the equation above to compute the annual fees, Campbell and Thompson (2008) demonstrated that the investor’s total realized mean-variance utility can be alternatively measured via the Sharpe ratio. That is, the annual fees can be computed using

\[
SR \left( R_{t,t+q}^{\text{mod}_1} \right) = SR \left( R_{t,t+q}^{\text{mod}_2} - q \times \Delta \right),
\]

where \( SR(\cdot) \) denotes the Sharpe ratio.

In our computations, we assume that the investor’s risk aversion \( \gamma = 5 \) (as in Kirby and Ostdiek, 2012). Since we do not have data for the real risk-free rate of return, to perform the computations, we assume that the nominal annual risk-free rate of return equals the annual inflation rate. Therefore, in real terms, \( r_{t,t+1}^{\text{free}} = 0 \). We measure the annual performance fees over our total OOS period 1921-2011. Table 6 reports the Sharpe ratios associated with each predictive model and the estimated annual fees measured in basis points. The results are reported for two values of \( q \): \( q = 1 \) and \( q = 15 \). In the first case, the investor rebalances his portfolio once a year; in the second case, the investor rebalances his portfolio once in 15 years.

First, we consider the case in which the investor rebalances his portfolio once a year. In this case, the Sharpe ratios of all predictive models, which perform statistically significantly better than the historical-mean model, are higher than the Sharpe ratio of the historical-mean model. The advantages of these predictive models translate into significant utility gains. Specifically, risk-averse investors would be willing to pay from 30 to 77 basis points in fees per year to switch...
from the historical-mean model to a model with superior forecast accuracy. In contrast to these models, our results indicate that the model that uses the long-term bond yield as a predictor exhibits inferior forecast accuracy compared with historical-mean model. As a result, not only is the Sharpe ratio of this model lower than that of the historical-mean model, but also the investor would need to be paid 20 basis points in fees per year to switch from the historical-mean model to the bond yield model.

When the investor can rebalance his portfolio once a year, the price-earnings model performs best while the mean-reverting model performs second best. However, when the investor decreases the portfolio revision frequency, the performance gains delivered by the price-earnings model diminish whereas the performance gains provided by the mean-reverting model remain rather stable. When the investor rebalances his portfolio once in 15 years, the performance gains of the price-earnings model virtually disappear. In contrast, the performance gains of the mean-reverting model (as measured in annual fees) remain virtually intact. Therefore, in cases in which the investor has to make long-term allocation decisions, the mean-reverting model delivers the highest performance gains.

V. SUMMARY AND CONCLUSIONS

We began the paper by performing two tests of the random walk hypothesis using the real Standard and Poor's Composite Stock Price Index data for the period from 1871 to 2011. In particular, we investigated the time series properties of the index returns at increasing horizons of up to 40 years. In our tests of the random walk hypothesis, we used two well-known test statistics: the autocorrelation of multi-year returns and the variance ratio. In the context of the null hypothesis, our goal was to test whether the index returns are distributed independently of their ordering in time. To estimate the significance level of the test statistics under the null hypothesis, we employed randomization methods that are free of distributional assumptions.

Rather surprisingly, given the seemingly insufficient span of available historical observations of the returns on the stock index, both test statistics allowed us to reject the random walk hypothesis at conventional significance levels over very long horizons of approximately 30-34 years. By studying the impact of the sample period on the test statistics, we concluded that mean reversion seems to be an extraordinarily strong phenomenon during the post-1926 period. Having performed the same randomization tests with stratification, we found that the results based on the use of the variance ratio are sensitive to the particular pattern of heteroskedasticity that occurred historically, while the results based on the use of the autocorrelation of multi-year returns are not.

Consequently, we do not have sufficiently strong evidence to claim that the variance ratio decreases with increasing investment horizons. In other words, our results cannot fully support the conventional belief that the stock market is safer for long-term investors. In contrast, we do have convincing evidence that suggests that a given change in price over 15-17 years tends to be reversed over the next 15-17 years by a predictable change in the opposite direction. Overall, our findings support the mean reversion hypothesis as the alternative to the random walk hypothesis. Our evidence of secular mean reversion in stock prices is robust to the choice of data source, deflator used to compute the real prices and returns, sample period, and test statistic.

The results of our tests provided evidence of in-sample predictability. However, the conventional wisdom states that in-sample evidence of stock return predictability might be a result of data mining. To guard against data mining, we investigated the performance of out-of-sample forecasts of multi-year returns. We demonstrated that the out-of-sample forecast provided by

A similar conclusion was drawn by Nelson and Kim (1993).
the mean-reverting model is statistically significantly better than the forecast provided by the historical-mean model. Moreover, the out-of-sample forecast accuracy of the mean-reverting model is comparable to that of Robert Shiller’s very popular (among practitioners) model that uses the cyclically adjusted price-earnings ratio as a predictor of long-horizon returns and to that of the model that uses the price-dividends ratio as a predictor of long-horizon returns. In addition, we demonstrated that the advantages of these three predictive models translate into significant utility gains. We found that in cases in which the investor has to make long-term allocation decisions, the mean-reverting model delivers the highest performance gains. Furthermore, in the post-1960 period, the mean-reverting model showed the best forecast accuracy among all competing models.

Given the main result of our study, it is natural to ask the following question. What caused this long-lasting mean reversion in stock market prices? In other words, what is the economic intuition behind this result? One possible answer is suggested by previous research on the link between demography and stock market returns and on the long-term variations in the birth rates and population growth in the US. In particular, on the one hand, Bakshi and Chen (1994), Dent (1998), Geanakoplos et al. (2004), and Arnott and Chaves (2012) observed an interrelationship between demography and US stock market returns and argued that demography determines stock market returns. On the other hand, the evidence presented by Kuznets (1958), Dent (1998), Berry (1999), and Geanakoplos et al. (2004) suggests the presence of secular trends in birth rates in the US that last from 10 to 20 years. Thus, if population growth exhibits long-term alternating periods of above-average and below-average rates, and demography is what determines stock market returns, then it is natural to expect that the stock market also exhibits long-term alternating periods of above-average and below-average returns.

A more elaborate model of cyclical dynamics of economic activity, interrelated with similar movements in other elements, was presented by Schlesinger (1949), Schlesinger (1986), Berry (1991), Berry et al. (1998), and Alexander (2004). These authors argued that the dynamics of economic activity in the US have a long-term rhythm (with a period of 12-18 years) of accelerated and retarded secular growth. This cyclical fluctuation in economic activity, in particular the alternation between long-term periods of good and bad economic times, gives rise to similar long-term fluctuations in social and political activities. In brief, a long-term period of rapid economic growth and technological development coincides with a conservative political wave (era). The conservative politics reduces the scope and the role of government in the life of the nation and frees up business and capital. Such a period is also characterized by higher population growth, increased inequality, and deflationary conditions. However, a long-term period of economic growth inevitably leads to a long-term stagflation crisis. During such a crisis, conservative leaders are replaced by liberal leaders committed to business regulation, social innovation, equity, and redistribution via an enhanced role of government. A liberal era is usually characterized by lower population growth, decreased inequality, and inflationary conditions. In our opinion, the secular mean-reverting behavior of the stock market fits nicely into this model of socioeconomic dynamics. It seems possible to demonstrate that conservative political waves are usually associated with above-average stock market returns, whereas during the liberal political waves, stock market returns are below average. The investigation of this topic is left for future research.

In the American two-party system, the Democratic Party traditionally adheres to the liberal policy, whereas the Republican Party maintains the conservative policy. Liberals believe in government action to achieve equal opportunity and equality for all. They advocate for government regulation of the economy, higher taxes (primarily on the wealthy), and redistribution of wealth. In contrast, conservatives believe in free markets, limited government, lower taxes, and individual liberty. There is strong evidence that in the US, stock market returns vary across political cycles; see Hensel and Ziemba (1995) and Santa-Clara and Valkanov (2003). Yet, there are many examples of a Republican president implementing liberal policy and a
Secular Mean Reversion and Long-Run Predictability of the Stock Market

REFERENCES

Alexander, M. A. (2000). *Stock Cycles: Why Stocks Won’t Beat Money Market Over the Next Twenty Years*. iUniverse Star.

Alexander, M. A. (2004). *Cycles in American Politics*. iUniverse Star.

Arnott, R. D. and Chaves, D. B. (2012). ‘Demographic Changes, Financial Markets, and the Economy’, *Financial Analysts Journal*, 68, pp. 23–46.

Asness, C. S. (2003). ‘Fight the fed model’, *Journal of Portfolio Management*, 30, pp. 11–24.

Bakshi, G. S. and Chen, Z. (1994). ‘Baby boom, population aging, and capital markets’, *Journal of Business*, 67, pp. 165–202.

Balvers, R., Yangru, W. and Gilliland, E. (2000). ‘Mean reversion across national stock markets and parametric contrarian investment strategies’, *Journal of Finance*, 55, pp. 745–72.

Berge, K., Consigli, G. and Ziemba, W. T. (2008). ‘The predictive ability of the bond-stock earnings yield differential model’, *Journal of Portfolio Management*, 34, pp. 53–80.

Berry, B. J. L. (1991). *Long-Wave Rhythms in Economic Development and Political Behavior*. Johns Hopkins University Press.

Berry, B. J. L. (1999). ‘Fertility cycles: a note on onset and periodicity’, *Population and Environment*, 21, pp. 149–54.

Berry, B. J. L., Elliot, E., Harpham, E. J. and Kim, H. (1998). *The Rhythms of American Politics*. University Press of America.

Bodie, Z., Kane, A. and Marcus, A. J. (2007). *Investments*. McGraw Hill.

Bossaerts, P. and Hillion, P. (1999). ‘Implementing statistical criteria to select return forecasting models: what do we learn?’, *Review of Financial Studies*, 12, pp. 405–28.

Campbell, J. Y. (1987). ‘Stock returns and the term structure’, *Journal of Financial Economics*, 18, pp. 373–99.

Campbell, J. Y. and Shiller, R. J. (1988). ‘The dividend-price ratio and expectations of future dividends and discount factors’, *Review of Financial Studies*, 1, pp. 195–227.

Campbell, J. Y. and Shiller, R. J. (1998). ‘Valuation ratios and the long-run stock market outlook’, *Journal of Portfolio Management*, 24, pp. 11–26.

Campbell, J. Y. and Shiller, R. J. (2001). ‘Valuation ratios and the long-run stock market outlook: an update’, Working paper, Harvard University and Yale University.

Campbell, J. Y. and Thompson, S. B. (2008). ‘Predicting excess stock returns out of sample: can anything beat the historical average?’, *The Review of Financial Studies*, 21, pp. 1509–31.

Campbell, J. Y. and Yogo, M. (2006). ‘Efficient tests of stock return predictability’, *Journal of Financial Economics*, 81, pp. 27–60.

Cavanagh, C. L., Elliott, G. and Stock, J. H. (1995). ‘Inference in models with nearly integrated regressors’, *Econometric Theory*, 11, pp. 1131–47.

Cochrane, J. H. (1988). ‘How big is the random walk in GNP?’, *Journal of Political Economy*, 96, pp. 893–920.

Dent, H. (1998). *The Roaring 2000s*. Simon and Shuster, New York.

Diebold, F. X. and Mariano, R. S. (1995). ‘Comparing predictive accuracy’, *Journal of Business and Economic Statistics*, 13, pp. 253–63.

Easterling, E. (2005). *Unexpected Returns: Understanding Secular Stock Market Cycles*. Cypress House.

Estrada, J. (2006). ‘The fed model: a note’, *Finance Research Letters*, 3, pp. 14–22.

Estrada, J. (2009). ‘The fed model: the bad, the worse, and the ugly’, *Quarterly Review of Economics and Finance*, 49, pp. 214–38.

Fama, E. F. (1970). ‘Efficient capital markets: a review of theory and empirical work’, *Journal of Finance*, 25, pp. 383–416.

Fama, E. F. (1981). ‘Stock returns, real activity, inflation, and money’, *American Economic Review*, 71, pp. 545–65.

Democratic president implementing conservative policy (Berry et al. (1998)). Therefore, there is no strong alignment between political cycles and conservative and liberal political waves.

© 2016 Board of Trustees of the Bulletin of Economic Research and John Wiley & Sons Ltd
Fama, E. F. (1990). ‘Term-structure forecasts of interest rates, inflation, and real returns’, *Journal of Monetary Economics*, 25, pp. 59–76.

Fama, E. F. and French, K. R. (1988a). ‘Dividend yields and expected stock returns’, *Journal of Financial Economy*, 22, pp. 3–25.

Fama, E. F. and French, K. R. (1988b). ‘Permanent and temporary components of stock Prices’, *Journal of Political Economy*, 96, pp. 246–73.

Fama, E. F. and French, K. R. (1989). ‘Business conditions and expected returns on stocks and bonds’, *Journal of Financial Economics*, 25, pp. 23–49.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Fleming, J., Kirby, C. and Ostdiek, B. (2001). ‘The economic value of volatility timing’, *Journal of Finance*, 56, pp. 329–52.

Geanakoplos, J., Magill, M. and Quinzii, M. (2004). ‘Demography and the long-run predictability of the stock market’, *Brookings Papers on Economic Activity*, 1, pp. 241–325.

Goyal, A. and Welch, I. (2003). ‘Predicting the equity premium with dividend ratios’, *Management Science*, 49, pp. 639–54.

Gropp, J. (2003). ‘Mean reversion of size-sorted portfolios and parametric contrarian strategies’, *Managerial Finance*, 29, pp. 5–21.

Gropp, J. (2004). ‘Mean reversion of industry stock returns in the U.S., 1926-1998’, *Journal of Empirical Finance*, 11, pp. 537–51.

Hensel, C. R. and Ziemba, W. T. (1995). ‘United States investment returns during democratic and republican administrations, 1928-1993’, *Financial Analysts Journal*, 51, pp. 61–9.

Hirsch, J. A. (2012). *The Little Book of Stock Market Cycles*. John Wiley and Sons, Joboken, NJ.

Jegadeesh, N. (1991). ‘Seasonality in stock price mean reversion: evidence from the U.S. and the U.K.’, *Journal of Finance*, 46, pp. 1427–44.

Katsenelson, V. N. (2007). *Active Value Investing: Making Money in Range-Bound Markets*. Wiley Finance.

Kim, M. J., Nelson, C. R. and Startz, R. (1991). ‘Mean reversion in stock prices? A reappraisal of the empirical evidence’, *Review of Economic Studies*, 58, pp. 515–28.

Kirby, C. and Ostdiek, B. (2012). ‘It’s all in the timing: simple active portfolio strategies that outperform naive diversification’, *Journal of Financial and Quantitative Analysis*, 47, pp. 437–67.

Koivu, M., Pennanen, T. and Ziemba, W. T. (2005). ‘Cointegration analysis of the fed Model’, *Finance Research Letters*, 2, pp. 248–59.

Kuznets, S. (1958). ‘Long swings in the growth of population and in related economic Variables’, *Proceedings of the American Philosophical Society*, 102, pp. 25–52.

Lander, J., Orphanides, A. and Douvogiannis, M. (1997). ‘Earnings forecasts and the predictability of stock returns: evidence from trading the S&P’, *Journal of Portfolio Management*, 23, pp. 24–35.

Lanne, M. (2002). ‘Testing the predictability of stock returns’, *Review of Economics and Statistics*, 84, pp. 407–15.

Leroy, S. F. (1982). ‘Expectations models of asset prices: a survey of theory’, *Journal of Finance*, 37, pp. 185–217.

Lewellen, J. (2004). ‘Predicting returns with financial ratios’, *Journal of Financial Economics*, 74, pp. 209–35.

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

Maio, P. (2013). ‘The fed model and the predictability of stock returns’, *Review of Finance*, 17, pp. 1489–533.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall/CRC.

Marriott, F. H. C. and Pope, J. A. (1954). ‘Bias in the estimation of autocorrelations’, *Biometrika*, 41, pp. 390–402.

© 2016 Board of Trustees of the Bulletin of Economic Research and John Wiley & Sons Ltd
Secular Mean Reversion and Long-Run Predictability of the Stock Market

McCracken, M. W. (2007). ‘Asymptotics for out of sample tests of granger causality’, *Journal of Econometrics*, 140, pp. 719–52.

Nelson, C. R. and Kim, M. J. (1993). ‘Predictable stock returns: the role of small sample bias’, *Journal of Finance*, 48, pp. 641–61.

Noreen, E. W. (1989). *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. John Wiley and Sons, New York.

Officer, R. R. (1973). ‘The variability of the market factor of the New York stock exchange’, *Journal of Business*, 46, pp. 434–53.

Orcutt, G. H. and Irwin, J. O. (1948). ‘A study of the autoregressive nature of the time series used for Tinbergen’s model of the economic system of the United States’, *Journal of the Royal Statistical Society. Series B*, 10, pp. 1–53.

Poterba, J. M. and Summers, L. H. (1988). ‘Mean reversion in stock prices: evidence and implications’, *Journal of Financial Economics*, 22, pp. 27–59.

Rapach, D. E. and Wohar, M. E. (2005). ‘Valuation ratios and long-horizon stock price Predictability’, *Journal of Applied Econometrics*, 20, pp. 327–44.

Richardson, M. (1993). ‘Temporary component of stock prices: a skeptic’s view’, *Journal of Business and Economic Statistics*, 11, pp. 199–207.

Richardson, M. and Stock, J. (1989). ‘Drawing Inferences From Statistics Based on Multi-Year Asset Returns’, *Journal of Financial Economics*, 25, pp. 323–48.

Rogers, J. (2005). *Hot Commodities: How Anyone Can Invest Profitably in the World’s Best Market*. Random House Publishing Group.

Santa-Clara, P. and Valkanov, R. (2003). ‘The presidential puzzle: political cycles and the stock market’, *Journal of Finance*, 58, pp. 1841–72.

Schlesinger, Jr., A. M. (1986). *The Cycles of American History*. Boston, MA: Houghton Mifflin.

Schlesinger, Sr., A. M. (1949). *Paths to the Present*. New York: Macmillan Co.

Schwert, G. W. (1989). ‘Why does stock market volatility change over time?’, *Journal of Finance*, 44, pp. 1115–53.

Shiller, R. J. (1989). *Market Volatility*. The MIT Press.

Shiller, R. J. (2000). *Irrational Exuberance*. Princeton University Press.

Siegel, J. (2002). *Stocks for the Long Run*. McGraw-Hill Companies.

Stambaugh, R. F. (1999). ‘Predictive regressions’, *Journal of Financial Economics*, 54, pp. 375–421.

Summers, L. H. (1986). ‘Does the stock market rationally reflect fundamental values?’, *Journal of Finance*, 41, pp. 591–601.

Torous, W., Valkanov, R., and Yan, S. (2004). ‘On predicting stock returns with nearly integrated explanatory variables’, *Journal of Business*, 77, pp. 937–66.

Welch, I. and Goyal, A. (2008). ‘A comprehensive look at the empirical performance of equity premium prediction’, *Review of Financial Studies*, 21, pp. 1455–508.