Equity Premium Forecasts Tend to Perform Worse Against a Buy-and-Hold Benchmark

Gunter Löffler*
Ulm University
gunter.loeffler@uni-ulm.de

August 24, 2021

Abstract

The economic gains from using equity premium forecasts are usually assessed by comparing a forecast-based strategy to a strategy based on the trailing historical mean. Whether these economic gains are statistically significant remains mostly untested. This paper shows that a buy-and-hold benchmark can be much harder to beat than the historical-mean benchmark and that the practice of not testing the statistical significance of economic gains can lead to questionable conclusions. The findings rest on an examination of many hypothetical sample periods and the replication of two widely cited papers (Rapach, Strauss, and Zhou 2010; Rapach, Ringgenberg, and Zhou 2016).

***JEL classification*:** G12, G17

***Keywords*:** equity premium; predictability; out-of-sample; utility gains

*I would like to thank an anonymous referee, Amit Goyal and Ivo Welch (the editor) for their comments and advice.

There is extensive academic research on the predictability of the equity premium. In 2008, Ivo Welch and Amit Goyal (2008) showed that a plethora of suggested prediction models did not reliably forecast the equity premium in an out-of-sample setting. Since then, the literature has proposed many new prediction models. Based on evaluation methods employed by Welch and Goyal (2008), researchers have concluded that the equity premium is predictable and that investors would have benefited from using the forecasts.

To assess economic gains, the extant literature proceeds as follows: Follow an investment strategy that determines an optimal equity weight with the help of a forecast and compare the strategy's performance to that of a strategy that optimizes based on the historical mean. The analysis is performed in a (pseudo) out-of-sample setting. With respect to data availability, the strategies could be implemented in real time.

Based on an examination of a large number of different sample periods, the present paper first shows that the historical-mean benchmark is usually less stringent than a buy-and-hold strategy, i.e., a 100% passive investment in the market index. I then replicate two widely cited post-2008 papers that predict the US equity premium. The economic gains reported by David E. Rapach, Jack K. Strauss, and Guofu Zhou (2010) more or less disappear when the buy-and-hold benchmark is used. Rapach, Matthew C. Ringgenberg, and Zhou (2016) is an example of a paper that reports performance advantages relative to the buy-and-benchmark but does not test their statistical significance. My replication shows that the performance advantages are insignificant, with p-values above 20%.

The remainder of the paper is structured as follows. Section 1 introduces the out-of-sample methodology and discusses aspects relevant in benchmark choice. Section 2 compares benchmark performance for various sample periods. Section 3 presents the results of the replication exercise, and section 4 concludes the paper.

1. Assessing and testing the economic value of equity premium predictions

The standard toolbox

In equity premium prediction, researchers use models to forecast the excess return that an investment in the aggregate stock market offers relative to a risk-free investment. Consider a prediction model that is built and tested using data from $t = 1$ to T, say January 1970 to December 2020. For an out-of-sample analysis, one would choose an evaluation period $t = m$ to T, say January 1990 to December 2020. For each month in this evaluation period, researchers would recursively generate forecasts using only information that is available at the time a forecast is made.

Whether a forecast model is economically useful is usually examined by comparing the risk-adjusted performance of a forecast-based investment strategy to that of a benchmark. The common practice is to examine the utility of an investor with a mean–variance objective. John Y. Campbell and Samuel B. Thompson (2008) use the objective

$$\mathrm{E}\big[R_p\big] - \frac{1}{2}\gamma\sigma^2\big(R_p\big), \tag{1}$$

where R_p is the return of the strategy portfolio, and γ is the investor's risk aversion. If an investor must decide how much to invest in equity (with return R) and how much to invest in the risk-free asset (with return R^f), the optimal solution for the equity weight is

$$w^* = \frac{1}{\gamma}\frac{\mathrm{E}[R] - R_f}{\sigma^2(R)}. \tag{2}$$

In order to determine the optimal investment for time t as part of an out-of-sample analysis, researchers choose a plausible value for γ, use a forecast $\hat{R}_t^e$ as an estimate for $E[R_p] - R_f$, and estimate $\sigma^2(R)$ with data prior to t. The underlying one-period optimization model

is silent on how to deal with a risk-free rate that varies over time. Researchers usually estimate the variance with excess returns.[1]

The average utility level from using forecasts of approach i can be estimated by replacing the moments in (1) with their sample counterparts

$$u_i = \frac{1}{T-m+1}\sum_{t=m}^{T} R_{Pit} - \frac{1}{2}\gamma\frac{1}{T-m}\sum_{t=m}^{T}\left(R_{Pit} - \frac{1}{T-m+1}\sum_{t=m}^{T} R_{Pit}\right)^2. \qquad (3)$$

The usual benchmark is based on the trailing mean computed with returns from the sample that was used to produce the forecasts of approach i. In the application of (2), this trailing mean is used as an estimate for $\mathrm{E}(R)$; other assumptions do not differ from the calculations for forecast approach i. The utility gain or certainty equivalent return (CER) is then

$$\Delta u = u_{forecast\ based} - u_{historical\ mean\ based}. \qquad (4)$$

Usually, Δu is expressed on a per annum basis. With a monthly return frequency, for example, one would multiply Δu by 12. The utility gain can then be interpreted as the annual management fee an investor would be willing to pay for getting access to the forecast approach.

The buy-and-hold benchmark and Sharpe ratio analysis

A less frequently used benchmark is the buy-and-hold strategy, which has a constant equity weight of one. A utility-based comparison puts the buy-and-hold strategy on a disadvantage because it may be too aggressive or too conservative for the risk aversion considered in the analysis. This can be addressed with the analysis of Sharpe ratios. Fixed-weight strategies that differ only in their fixed weight have the same Sharpe ratio. Effectively, comparing an optimizing strategy with the buy-and-hold portfolio therefore

[1] For a regression-based forecast, one could also use the predictive variance (cf. Shmuel Kandel and Robert F. Stambaugh, 1996) or the residual variance of the forecast regression. Given that the R^2 of predictive regressions is typically small, the choice may not matter much in practice.

does not assume that investors are fully invested in equity; it assumes that they allocate a constant fraction of their wealth to equity and the remainder to the risk-free asset. On the other hand, Sharpe ratios are also an appropriate criterion for strategies that optimize according to (1) because mean-variance optimization with a risk-free asset results in a maximization of the Sharpe ratio.

Statistical significance of economic gains

As noted by Michael W. McCracken and Giorgio Valente (2018), researchers usually do not test the statistical significance of economic performance advantages. There are several approaches for statistical tests of utility gains. The first test I perform is the one suggested by Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal (2009, footnote 18), which uses the delta method to derive the distribution of Δu. The second test builds on realized utilities. Let V_t be an estimate of realized equity volatility in period $[t-1, t]$. Realized utility is then

$$u_{it} = R_{Pit} - \frac{1}{2}\gamma\, w_i^2 V_t^2, \tag{5}$$

and overall utility can be estimated as the mean of period-by-period realized utilities:

$$u_i = \frac{1}{T-m+1}\sum_{t=m}^{T} u_{it}. \tag{6}$$

Because equation (5) provides a time series of utilities, a straightforward procedure is to conduct a t-test on the time series of utility differences. The inspection of realized utilities was proposed by Wessel Marquering and Marno Verbeek (2004), and a corresponding test was used by Thomas Dangl and Michael Halling (2012). I will use the sum of squared daily market returns as an estimate of V_t^2.

For the purpose of this paper, I choose these two tests rather than the ones suggested by McCracken and Valente (2018) because the latter do not allow for recursive detrending of

predictors, which is applied in one of the papers that I replicate (Rapach, Ringgenberg, and Zhou, 2016). To test the statistical significance of Sharpe ratio differences, I use the procedure of J. D. Jobson and Bob M. Korkie (1981) and Christoph Memmel (2003).

Benchmark choice—a discussion

Having introduced the methodology, it is worth looking at aspects that are relevant when choosing a benchmark for equity premium forecasts. The optimizing historical-mean benchmark uses the same framework as the forecast-based strategy. One would therefore expect that performance differences can more easily be attributed to the quality of the studied equity forecast. However, confounding factors can still be at work. Given that papers in the predictability literature usually constrain the equity weights to some interval (e.g., 0% to 150%), an upward-biased equity premium forecast will bring the forecast-based optimized strategy closer to a fixed-weight strategy because it will more frequently make the upper allocation constraint binding. If a fixed-weight strategy happens to perform well with the data used in a study, a bias in equity premium forecasts could result in the forecast-based strategy outperforming the historical-mean benchmark.[2] Another possible confounder is variance. If the variance estimate used for determining optimal weights in (2) is not the best available estimator, a forecast that negatively correlates with errors in the variance estimates will lead to better performance because it neutralizes errors in the variance estimate.

The buy-and-hold strategy is easy to implement. It is also naïve in the sense that it does not optimize. However, given the mixed evidence on the performance of optimized

[2] To illustrate the possible magnitude of this confounding factor, I analyzed a hypothetical forecast for the data and setting used in Rapach, Ringgenberg, and Zhou (2016). A biased monthly forecast, constructed as a constant c plus the historical mean, leads to a utility gain of 1.39% relative to the historical-mean benchmark whenever c is 0.81% per month or higher.

portfolios,[3] it is possible that a naïve strategy could be superior to optimized portfolios. A rational investor should consider a buy-and-hold strategy as an option if the data show sufficiently strong evidence that its risk-adjusted performance is better than that of the optimizing strategy the investor planned to pursue.

Still, the buy-and-hold benchmark can be viewed as ad hoc. Investors questioning the empirical robustness of optimizing strategies could also consider other strategies. If performance is compared using Sharpe ratios, however, the buy-and-hold strategy loses some of this arbitrariness because it will lead to the same Sharpe ratio as other fixed-weight allocation strategies.

In addition, the optimizing strategies studied in the literature also have an ad-hoc component in that they usually constrain the equity weights to some interval. Campbell and Thompson (2008, p.1525) motivate this constraint by mentioning that it makes allocations realistic. With the same motivation, however, and supported by the fairly narrow allocation bands of large investment vehicles, one could also constrain the equity weight to between 40% to 100% or to between 50% to 75%.[4] Viewed in this light, a fixed-weight allocation strategy is a limiting case of the optimizing benchmark, which arises when the chosen range of admissible equity weights shrinks to zero. The practical difference remains, of course, which could mean that the buy-and-hold benchmark will set the bar too low because it does not optimize or take risk aversion into account. This

[3] Cf. DeMiguel, Garlappi, and Uppal (2009).

[4] One of the largest US balanced mutual funds (American Balanced Fund) describes its policy in its prospectus as follows: "Normally the fund will maintain at least 50% of the value of its assets in common stocks and at least 25% of the value of its assets in debt securities, including money market securities." In the period from 2003 to 2020, the actual equity weight at reporting dates lay in an even narrower interval, between 55% and 74%. See Internet Appendix Figure B4 for details and sources. The prospectus of another large US balanced fund (Vanguard Wellington Fund) describes an equity target range that is narrower still: "The Fund invests 60% to 70% of its assets in dividend-paying and, to a lesser extent, non-dividend-paying common stocks of established large companies."

constitutes a strong argument against the sole use of the buy-and-hold benchmark, but it does not invalidate its use as an additional benchmark. If a forecast model does not cross a bar that is too low, the economic significance of performance gains becomes doubtful.

Another aspect one could consider is the viability of strategy choices from the perspective of investors. If investors contemplate a fixed-weight strategy, it would not be obvious to them which weight they should choose. However, the same holds for the optimizing strategy because without some information collection and analysis investors would rarely know what their risk aversion γ is, how to best estimate the return variance, and how to best estimate the unconditional mean return. In such a situation, many investors are likely to turn to financial advisers. The fact that fixed-weight allocations are frequently recommended suggests that they are a realistic alternative to optimizing benchmarks.[5]

A study of actual portfolio policies is also likely to lead to different conclusions regarding which benchmark is more relevant. For investors who use optimization to determine their investment strategy, the optimizing benchmark appears suitable. For institutional investors with a fixed-allocation benchmark, the buy-and-hold strategy could appear to be a sensible choice, particularly when the performance comparison is based on Sharpe ratios.

Another question could also be asked: Which benchmark comes closest to the position of the average investor? Average portfolio weights will be equal to the market weights, which change from day to day with price fluctuations and security issuance. Hence, average weights will not be constant, but they may be better approximated by a fixed-allocation strategy than by a strategy whose weights fluctuate more widely.[6]

[5] An example of a fixed-weight allocation recommendation is the "Select a plan and stick with it" recommendation given by Charles Schwab; see https://www.schwabmoneywise.com/public/moneywise/essentials/finding_the_right_asset_allocation

[6] The weights of the optimizing historical-mean benchmark in the out-of-sample periods of the two studies that I replicate fluctuate between 0.70 and 1.50 (Rapach, Strauss, and Zhou, 2010) and 0.36 and 1.50 (Rapach, Ringgenberg, and Zhou, 2016). See Internet Appendix Figure B5 for details.

To conclude, there appear to be valid arguments for both benchmarks, and it seems reasonable to use both.

2. Attractiveness of the buy-and-hold benchmark

For typical parameter choices and a wide range of different sample periods, this section examines how the buy-and-hold strategy fares compared to the commonly used benchmark that optimizes the equity weight based on the trailing historical mean. Following many other papers, my analysis employs the data set compiled and provided by Amit Goyal,[7] from which I take the equity index data for the S&P 500 as well as the risk-free rate.

To determine optimal equity weights with equation (2), I use the assumptions from Campbell and Thompson (2008): I set the risk aversion γ to 3, estimate the variance of excess returns with five years of monthly data, and constrain the optimal portfolio weights to between 0 and 1.5.

I consider all 50-year samples that can be constructed with data from 1926:01 to 2020:12. In each sample, I let the out-of-sample period start 20 years after the sample begins so that the out-of-sample period has a length of 30 years. The trailing mean that is used for optimizing weights is the trailing mean from the respective sample. This corresponds to the usual practice in the literature: A study using a predictor that is available only, for example, from 1965 on would estimate the historical mean return with data starting in 1965, even though a larger return history is available.

Figure 1 shows the utility gains as well as the Sharpe ratio gains that investors would have achieved if they had followed the buy-and-hold-strategy instead of the strategy that optimizes the equity weight with equation (2). In 84.1% of all samples, the buy-and-hold

[7] http://www.hec.unil.ch/agoyal/. Details on the data are given in Welch and Goyal (2008).

strategy leads to a utility gain; in 92.2% of all samples, it leads to a higher Sharpe ratio. Importantly, the empirical attractiveness of the buy-and-hold strategy is not only evident ex post with today's knowledge. Much of the evidence presented in Figure 1 would have been available by the year 2008, when the publication of Welch and Goyal (2008) sparked a wave of new prediction models.

[Insert Figure 1 here]

To assess the stability of results, I consider several variations. With a risk aversion γ of five, the buy-and-hold portfolio continues to lead to a better performance in more than three quarters of all samples. The same holds if the initial estimation period in the 50-year samples is set to 30 years instead of 20 years, or if the variance used to determine optimal weights is taken to be constant.[8] The latter variation is not common in the predictability literature but it is worthwhile exploring because the variance might act as a confounder in the comparison between the buy-and-hold strategy and optimizing strategies. Detailed results for the three variations are presented in the Internet Appendix, Figures B1 to B3.

3. Impact of benchmark choice on published results

This section illustrates the consequences of benchmark choice for two papers, which I select as follows. To identify the sample of papers from which I choose, I consider papers that—according to the Social Science Citation Index—cite Campbell and Thompson (2008), the seminal paper on the out-of-sample testing approach used in the literature. Then, I rank these papers according to their own citations, again using information from the Social Science Citation Index. I first select the mostly highly cited paper, which is Rapach, Strauss, and Zhou (2010). I then go down the list of citation-ranked papers until I

[8] I set the variance to 0.1875%, corresponding to an annual volatility of 15%.

find a paper that shows results for the buy-and-hold benchmark, which leads me to Rapach, Ringgenberg, and Zhou (2016).

In my replication analysis, I rely not only on the descriptions in the paper but also on code made available by the authors. Some implementation details that may be helpful for readers doing a replication of their own are listed in Internet Appendix A.

Replication of Rapach, Strauss, and Zhou (2010)

Rapach, Strauss, and Zhou (2010) predict quarterly excess returns of the S&P 500. Their sample period is 1947:1 to 2005:4, and the out-of-sample period starts in 1965:1. For 15 predictors—also examined in Welch and Goyal (2008)—Rapach, Strauss, and Zhou (2010) run individual predictive regressions. The forecasts from these individual regressions do not improve out-of-sample accuracy in terms of mean-squared prediction error, but combinations do. The most accurate combination forecast is the arithmetic mean of the 15 individual forecasts. In terms of utility gains, the best-performing model is the one that combines the individual forecasts based on their trailing, discounted mean square prediction error (DMSPE), using a discount factor θ of 0.9. Therefore, I replicate the analysis for the mean combination forecast as well as for the DMSPE (θ=0.9) combination forecast.

Table 1 shows the results. The values obtained through the replication do not exactly match, but they are quite close. One reason for this could be differences in the data. Like Rapach, Strauss, and Zhou (2010), I am using data provided by Amit Goyal, but the 2019 version of the file I am using contains differences to previous versions.

[Insert Table 1 here]

Different from Rapach, Strauss, and Zhou (2010), I also test the statistical significance of the economic gains. With the optimized benchmark used in the original paper, the majority of p-values are below 5%; none of the p-values are above 10%. When I then use the alternative buy-and-hold benchmark, economic gains drop substantially, and they are no longer significant, with the majority of p-values above 85%. Utility gains decline from 2.39% and 2.64% to –0.14% and 0.11%; Sharpe ratio differences decline from 0.090 and 0.102 to 0.002 and 0.014.[9]

The results therefore indicate that the buy-and-hold benchmark can be much more difficult to beat than the optimizing strategy that the literature favors as a benchmark.

Replication of Rapach, Ringgenberg, and Zhou (2016)

Rapach, Ringgenberg, and Zhou (2016) predict excess returns on the S&P 500 with a short interest index (SII), which has been made available by the researchers. For the replication of the out-of-sample analysis, one only needs to implement a recursive detrending of short interest activity, as described in the original paper. The sample period starts in 1973:01; the out-of-sample period extends from 1990:01 to 2014:12. For the replication, I use my own code but follow the code published by the authors; I also take the return data from the file made available by the authors.[10] To display the authors' results for differences in economic gains with high precision, I take values saved in the authors' result file.

Table 2 presents the results for one-month ahead forecasts; results for other horizons are summarized below. The replication is exact. With the benchmark that optimizes with the trailing historical mean, each performance metric is significant at a level better than 5%.

[9] As in the sensitivity analysis for Figure 1, I also consider the following variations: (i) set γ to 5 instead of 3, and (ii) determine optimal weights with a constant variance corresponding to an annual volatility of 15%. None of these changes leads to an economic gain relative to the buy-and-hold strategy that is significant on a level of 10%.

[10] http://apps.olin.wustl.edu/faculty/zhou/Returns_econ_tech_data_programs.zip

With the buy-and-hold benchmark, none of the performance metrics are significant at a level better than 20%.[11]

[Insert Table 2 here]

The replication thus illustrates the possible consequences of not conducting statistical tests of economic gains. Rapach, Ringgenberg, and Zhou (2016, p. 58) offer the following conclusion regarding the buy-and-hold benchmark: "(…) a buy-and-hold portfolio that passively holds the market portfolio produces CER gains well below those of SII, so that SII also easily outperforms a buy-and-hold strategy."

Terms such as "easily" can be used with different meanings, and the precise conclusions one should draw from statistical tests are not always obvious. Even so, it seems likely that p-values above 20% would not lead every reader to agree with the conclusion that the buy-and-hold strategy is "easily" outperformed.

A possible argument against the usefulness of statistical tests is that their power could be low. As evident from Table 2, however, the tests used here do lead to rejections of the null for a different, but weaker benchmark; as shown above, the same is true for the replication of Rapach, Strauss, and Zhou (2010). Power could be an issue, but its relevance is not obvious. The failure to reject the null that the outperformance relative to the buy-and-hold portfolio is zero cannot simply be explained by stating that the tests are incapable of identifying performance differences in such situations.

Rapach, Ringgenberg, and Zhou (2016) also present results for three-month, six-month, and 12-month ahead forecasts. The results do not lead to a change of conclusions. The economic gains relative to the buy-and-hold benchmark are never significant at a level better than 10%. Detailed results are presented in the Internet Appendix, Table B1.

[11] As in the sensitivity analysis for Figure 1, I also consider the following variations: (i) set γ to 5 instead of 3, and (ii) determine optimal weights with a constant variance corresponding to an annual volatility of 15%. None of these changes leads to an economic gain relative to the buy-and-hold strategy that is significant on a level of 10%.

Assessing hindsight bias

The relevance of the presented results could be questioned by pointing out that it will almost always be possible to find a benchmark that turned out to be more stringent than the one favored in some paper. The analysis of Figure 1 mitigates such concerns because the stringency of the buy-and-hold benchmark was visible at the time the two papers were written. Additional analysis shows that the performance advantage of the buy-and-hold strategy would also have been visible to the imaginary investors of the out-of-sample exercises in the two replicated papers. In the first ten years of the out-of-sample periods, the buy-and-hold portfolio already leads to higher utilities and higher Sharpe ratios than the strategy optimizing with the historical mean, and these patterns continue until the end of the out-of-sample periods. Detailed results are shown in the Internet Appendix (Figure B6).

4. Concluding remarks

Welch and Goyal (2008) have cast doubt on the robustness of equity premium prediction models published before 2008. In this paper, I examined two highly cited papers that were published after 2008. The results suggest that the benchmark favored in the literature, which is based on the historical mean computed with the data used for the prediction model, is not particularly stringent. When the buy-and-hold strategy is used as a benchmark, utility and Sharpe ratio gains drop by sizeable amounts, and the statistical significance of the performance advantages disappears. The results also illustrate that a common practice in the predictability literature should be examined. Researchers usually do not test whether economic gains from following a forecast are statistically significantly different from zero. As it turns out, even gains that appear large may not withstand tests that have been suggested in the literature.

References

Campbell, John Y. and Samuel B. Thompson, 2008, Predicting excess stock returns out of sample: Can anything beat the historical average? The Review of Financial Studies 21, 1509-1531.

Dangl, Thomas and Michael Halling, 2012, Predictive regressions with time-varying coefficients. Journal of Financial Economics 106, 157-181.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal, 2009, Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? The Review of Financial Studies 22, 1915-1953.

Jobson, J. D. and Bob M. Korkie, 1981, Performance hypothesis testing with the Sharpe and Treynor measures. The Journal of Finance 36, 889-908.

Kandel, Shmuel and Robert F. Stambaugh, 1996, On the predictability of stock returns: an asset-allocation perspective. The Journal of Finance 51, 385-424.

Marquering, Wessel and Marno Verbeek, 2004, The economic value of predicting stock index returns and volatility. Journal of Financial and Quantitative Analysis 39, 407-429.

McCracken, Michael W. and Giorgio Valente, 2018. Asymptotic inference for performance fees and the predictability of asset returns. Journal of Business & Economic Statistics 36, 426-437.

Memmel, Christoph, 2003, Performance hypothesis testing with the Sharpe ratio. Finance Letters 1, 21-23.

Rapach, David E., Matthew C. Ringgenberg, and Guofu. Zhou, 2016, Short interest and aggregate stock returns. Journal of Financial Economics 121, 46-65.

Rapach, David E., Jack K. Strauss, and Guofu Zhou 2010, Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. The Review of Financial Studies 3, 821-862.

Welch, Ivo and Amit Goyal, 2008, A comprehensive look at the empirical performance of equity premium prediction. The Review of Financial Studies 21, 1455-1508.

Table 1: Revisiting the performance of combination forecasts of the US equity premium from Rapach, Strauss, and Zhou (2010).

The results are based on a replication of the analysis of Rapach, Strauss, and Zhou (2010). The mean combination forecast is the arithmetic average of 15 bivariate forecast models. The DMSPE combines forecasts from the same 15 models using discounted prediction error weights. Utility gains Δu indicate the per annum gain of a mean-variance investor who switches from a benchmark strategy to one that optimizes the S&P 500 weight based on the combination forecast; Sharpe ratio differences are determined for the same set of strategies and benchmarks. Benchmarks are based on (i) the trailing mean computed from the model sample beginning in 1947 (as in the original paper) and (ii) a buy-and-hold strategy. P-values are two-sided. The out-of-sample period in the paper and the replication is 1965:1 to 2005:4.

Interpretation: Benchmark choice matters for Rapach, Strauss, and Zhou (2010). Utility gains reported in the original paper more or less disappear if the benchmark is taken to be the buy-and-hold portfolio.

Benchmark:	Optimized with Trailing Mean		Buy-and-Hold	
	Original Paper	Replication	Original Paper	Replication
Panel A: Results for mean combination forecast				
Utility Gain Δu (in %)	2.34	2.39	-	–0.14
p from delta method test	-	(0.012)	-	(0.851)
p from t-test on realized Δu	-	(0.046)	-	(0.953)
ΔSharpe Ratio	-	0.090		0.002
p from Jobson–Korkie test	-	(0.038)		(0.967)
Panel B: Results for DMSPE ($\theta = 0.9$) combination forecast				
Utility Gain Δu (in %)	2.59	2.64	-	0.11
p from delta method	-	(0.022)	-	(0.905)
p from t-test on realized Δu	-	(0.090)	-	(0.896)
ΔSharpe Ratio	-	0.102		0.014
p from Jobson–Korkie test	-	(0.064)		(0.791)

Table 2: Revisiting the performance of a forecast of the US equity premium from Rapach, Ringgenberg, and Zhou (2016).

The results are based on a replication of the analysis performed in Rapach, Ringgenberg, and Zhou (2016). The forecast is a regression-based prediction with a short interest index. Utility gains Δu indicate the per annum gain of a mean-variance investor who switches from a benchmark strategy to one that optimizes the S&P 500 weight based on the combination forecast; Sharpe ratio differences are determined for the same set of strategies and benchmarks. Benchmarks are (i) the trailing mean computed from the model sample beginning in 1973 (as in the original paper) and (ii) a buy-and-hold strategy. P-values are two-sided. The out-of-sample period in the paper and the replication is 1990:01 to 2014:12.

Interpretation: With the buy-and-hold benchmark, the economic gains from using the forecast model are not statistically significant. Testing the statistical significance of economic gains can cast a different light on the results. Performance differences that may appear large can lack statistical significance.

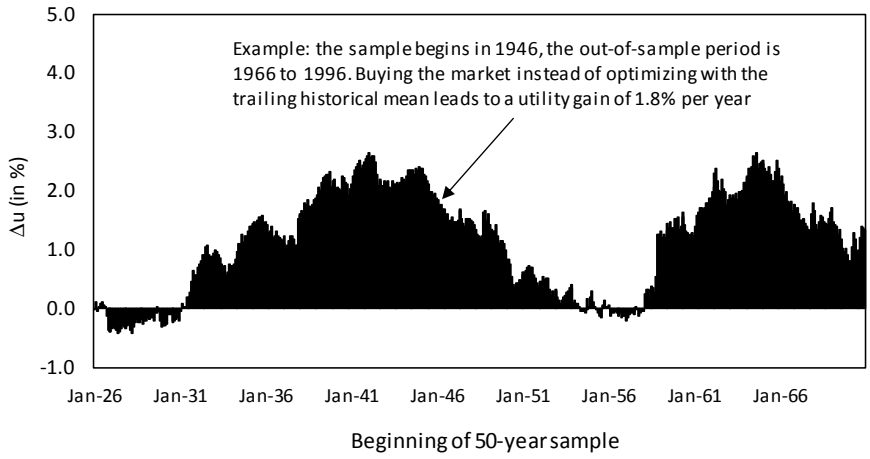
Benchmark:	Optimized with Trailing Mean		Buy-and-Hold	
	Original paper	Replication	Original Paper	Replication
Utility Gain Δu (in %)	4.17	4.17	2.46	2.46
p from delta method	-	(0.020)	-	(0.235)
p from t-test on realized Δu	-	(0.049)	-	(0.257)
ΔSharpe Ratio	0.270	0.270	0.150	0.150
p from Jobson–Korkie test		(0.023)		(0.268)

Figure 1: Performance of the buy-and-hold strategy relative to an optimizing strategy based on the historical mean—for different 50-year sample periods.

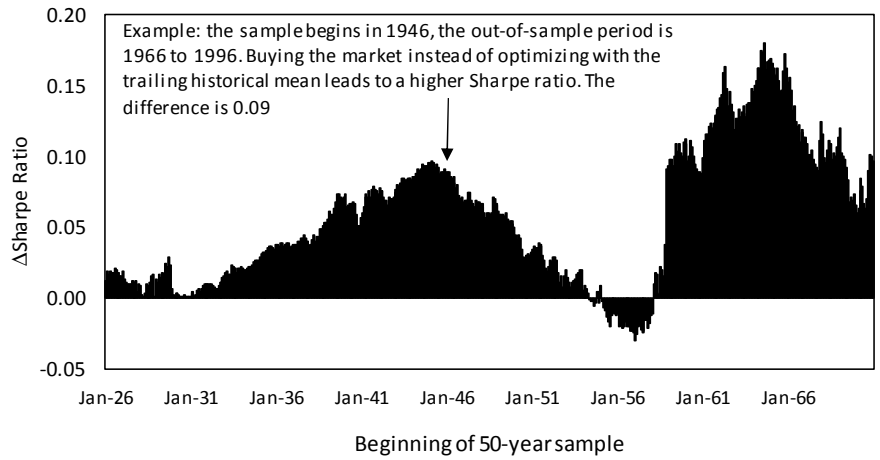
The graphs show economic gains when switching from a strategy that optimizes the S&P 500 weight based on the trailing historical mean to the buy-and-hold strategy. Utility gains Δu indicate the per annum gain of a mean-variance investor who switches from the optimizing strategy to the buy-and-hold strategy; Sharpe ratio differences are determined for the same switch. The analysis is conducted for different sample periods, each having a length of 50 years. The out-of-sample estimation period starts 20 years after the respective sample begins and lasts 30 years. The return frequency is monthly. Risk aversion is set to three, the variance of excess returns needed for the optimization is estimated with a five-year rolling window, and optimal equity weights are constrained to between 0% and 150%.

Interpretation: In the majority of the considered samples, the buy-and-hold strategy would have led to a better risk-adjusted performance than the optimizing strategy. Note that the performance differences are not cumulative. Each column represents the result for one 50-year sample.

Panel A: Δu = Utility from buy and hold minus utility from optimizing strategy



Panel B: ΔSharpe ratio = Sharpe ratio of buy and hold minus Sharpe ratio of optimizing strategy



Internet Appendix to "Equity Premium Forecasts Tend to Perform Worse Against a Buy-and-Hold Benchmark"

Appendix A: Implementation details for Rapach, Strauss, and Zhou (2010)

Rapach, Strauss, and Zhou (2010) predict logarithmic excess returns for the computation of forecast errors, but they do not explicitly describe which forecasts they use for the utility analysis. In their formulas for the forecast error metrics and the utility gains, the authors use the same symbol for predicted returns ($\hat{r}_t$).

Nevertheless, because the argument of the utility function is simple rather than logarithmic returns, I determine optimal portfolio weights with a combination forecast and means that are based on simple rather than logarithmic excess returns. This brings the utility gains in my replication closer to the ones reported in the paper.

Two coauthors of Rapach, Strauss, and Zhou (2010) have published code[12] that implements combination forecasts for another paper (Rapach and Zhou (2013))[13]. In this code, the trailing variance used in the computation of portfolio weights is estimated with the population variance (i.e., with division by *T*):[14]

```
FC_VOL(t)=mean(Y(R+P_0+(t-1)-window_VOL+1:R+P_0+(t-1)).^2)-...
     (mean(Y(R+P_0+(t-1)-window_VOL+1:R+P_0+(t-1))))^2;
```

whereas the variance of realized portfolio returns is determined with division by $T-1$:[15]

```
avg_utility=mean(return_portfolio)-0.5*gamma_MV*(std(return_portfolio))^2;
```

I follow this choice because it brings the utility gains in the replication closer to the ones reported in the paper that I replicate. The conclusions of the present paper do not change if division by $T-1$ is used for both variances.

[12] http://apps.olin.wustl.edu/faculty/zhou/HEF_2013_data_programs.zip

[13] Rapach, D. and G. Zhou, 2013, Forecasting stock returns. In Handbook of Economic Forecasting, Elsevier, Vol. 2, 328-383.

[14] From file: *Forecasts_quarterly.m*

[15] From file: *Perform_asset_allocation.m*

Appendix B

Table B1: Revisiting the performance of three- to twelve-month ahead forecasts of the US equity premium from Rapach, Ringgenberg, and Zhou (2016)

The table extends the analysis from Table 2 to the three-month, six-month, and 12-month horizons. The results are based on a replication of the analysis performed in Rapach, Ringgenberg, and Zhou (2016). The forecast is a regression-based prediction with a short interest index. Utility gains Δu indicate the per annum gain of a mean-variance investor who switches from a benchmark strategy to one that optimizes the S&P 500 weight based on the combination forecast; Sharpe ratio differences are determined for the same set of strategies and benchmarks. Benchmarks are based on (i) the trailing mean computed from the model sample beginning in 1947 (as in the original paper) and (ii) a buy-and-hold strategy. P-values are two-sided. The out-of-sample period in the paper and the replication is 1990:01 to 2014:12.

Interpretation: With the alternative benchmark, none of the performance metrics are statistically significant. Testing the statistical significance in utility gains or Sharpe ratios can cast a different light on results. Performance differences that may appear large can lack statistical significance.

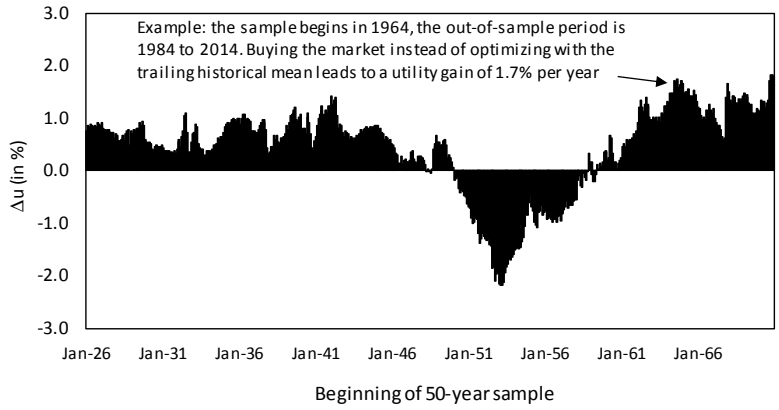
Benchmark:	Optimized with Trailing Mean		Buy-and-Hold	
	Original paper	Replication	Original Paper	Replication
	Panel A: Three-month forecast horizon			
Utility Gain Δu (in %)	4.65	4.65	2.06	2.06
p from delta method test	-	(0.023)	-	(0.380)
p from t-test on realized Δu	-	(0.046)	-	(0.389)
ΔSharpe Ratio	0.281	0.281	0.118	0.118
p from Jobson–Korkie test	-	(0.023)	-	(0.404)
	Panel B: Six-month forecast horizon			
Utility Gain Δu (in %)	5.44	5.44	3.18	3.18
p from delta method	-	(0.015)	-	(0.168)
p from t-test on realized Δu	-	(0.038)	-	(0.294)
ΔSharpe Ratio	0.353	0.353	0.199	0.199
p from Jobson–Korkie test	-	(0.027)		(0.217)
	Panel C: Twelve-month forecast horizon			
Utility Gain Δu (in %)	3.43	3.43	1.39	1.39
p from delta method	-	(0.139)	-	(0.635)
p from t-test on realized Δu	-	(0.115)	-	(0.491)
ΔSharpe Ratio	0.179	0.179	0.077	0.077
p from Jobson–Korkie test	-	(0.121)	-	(0.600)

Figure B1: Performance of the buy-and-hold strategy relative to an optimizing strategy based on the historical mean—for different 50-year sample periods and a risk aversion of five

The graphs show economic gains when switching from a strategy that optimizes the S&P 500 weight based on the historical mean to the buy-and-hold strategy. Utility gains Δu indicate the per annum gain of a mean-variance investor who switches from the optimizing strategy to the buy-and-hold strategy; Sharpe ratio differences are determined for the same switch. The analysis is conducted for different sample periods, each having a length of 50 years. The out-of-sample estimation period starts 20 years after the respective sample begins and lasts 30 years. The return frequency is monthly. Risk aversion is set to five, the variance of excess returns needed for the optimization is estimated with a five-year rolling window, and optimal equity weights are constrained to between 0% and 150%.

Interpretation: In the majority of the considered samples, the buy-and-hold strategy would have led to a better risk-adjusted performance than the optimizing strategy. Compared to the results in Figure 1, utility gains are lower because the buy-and-hold portfolio is too aggressive for a risk aversion of five; Sharpe ratio differences are slightly different from the ones in Figure 1 because the probability that the equity weight constraint becomes binding varies with the risk aversion.

Panel A: Δu = Utility from buy and hold minus utility from optimizing strategy



Panel B: ΔSharpe ratio = Sharpe ratio of buy and hold minus Sharpe ratio of optimizing strategy

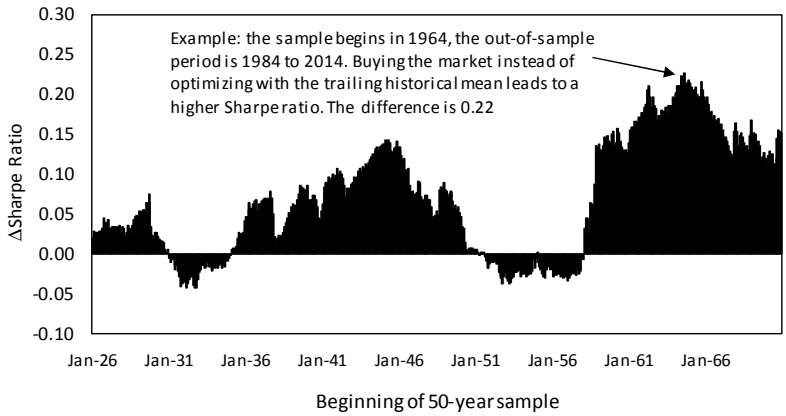
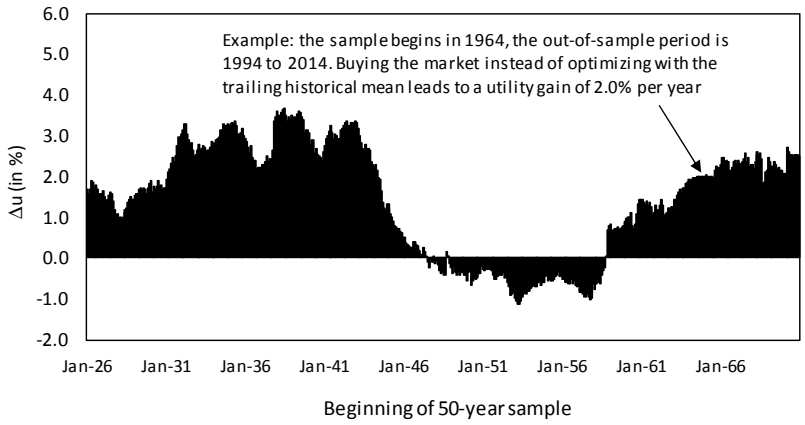


Figure B2: Performance of the buy-and-hold strategy relative to an optimizing strategy based on the historical mean—for different 50-year sample periods and an initial estimation period of 30 years

The graphs show economic gains when switching from a strategy that optimizes the S&P 500 weight based on the historical mean to the buy-and-hold strategy. Utility gains Δu indicate the per annum gain of a mean-variance investor who switches from the optimizing strategy to the buy-and-hold strategy; Sharpe ratio differences are determined for the same switch. The analysis is conducted for different sample periods, each having a length of 50 years. The out-of-sample estimation period starts 30 years after the respective sample begins and lasts 20 years. The return frequency is monthly. Risk aversion is set to three, the variance of excess returns needed for the optimization is estimated with a five-year rolling window, and optimal equity weights are constrained to between 0% and 150%.

Interpretation: In the majority of the considered samples, the buy-and-hold strategy would have led to a better risk-adjusted performance than the optimizing strategy.

Panel A: Δu = Utility from buy and hold minus utility from optimizing strategy



Panel B: ΔSharpe ratio = Sharpe ratio of buy and hold minus Sharpe ratio of optimizing strategy

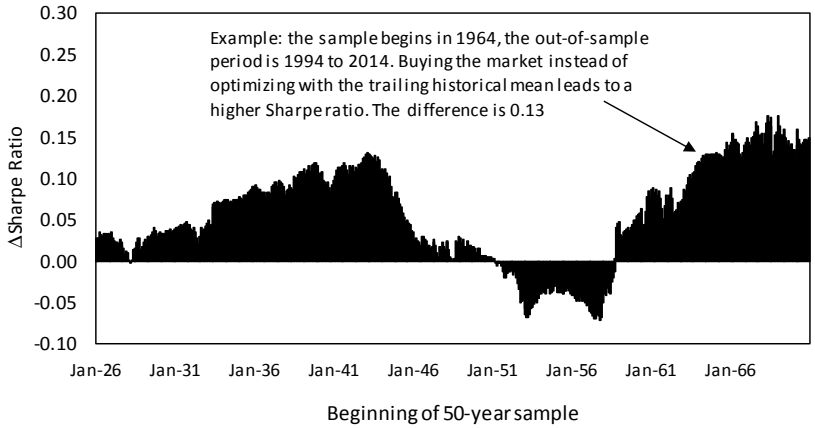
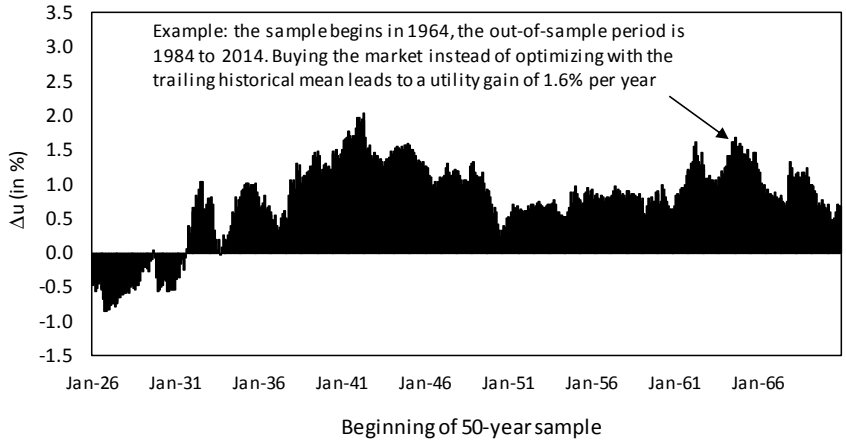


Figure B3: Performance of the buy-and-hold strategy relative to an optimizing strategy based on the historical mean—for different 50-year sample periods and with variance set to a constant

The graphs show economic gains when switching from a strategy that optimizes the S&P 500 weight based on the historical mean to the buy-and-hold strategy. Utility gains Δu indicate the per annum gain of a mean-variance investor who switches from the optimizing strategy to the buy-and-hold strategy; Sharpe ratio differences are determined for the same switch. The analysis is conducted for different sample periods, each having a length of 50 years. The out-of-sample estimation period starts 20 years after the respective sample begins and lasts 30 years. The return frequency is monthly. Risk aversion is set to three, optimal equity weights are constrained to between 0% and 150%, and the variance of excess returns needed for the optimization is set to 0.1875%, corresponding to an annual volatility of 15%.

Interpretation: In the majority of the considered samples, the buy-and-hold strategy would have led to a better risk-adjusted performance than the optimizing strategy.

Panel A: Δu = Utility from buy and hold minus utility from optimizing strategy



Panel B: ΔSharpe ratio = Sharpe ratio of buy and hold minus Sharpe ratio of optimizing strategy

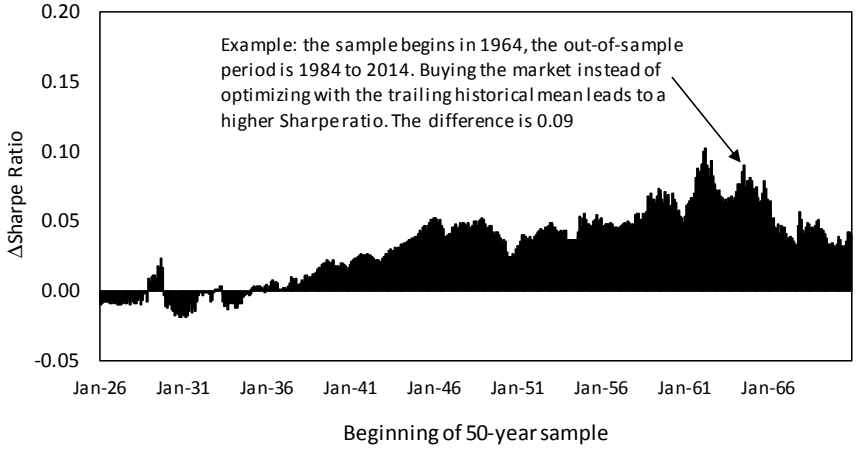
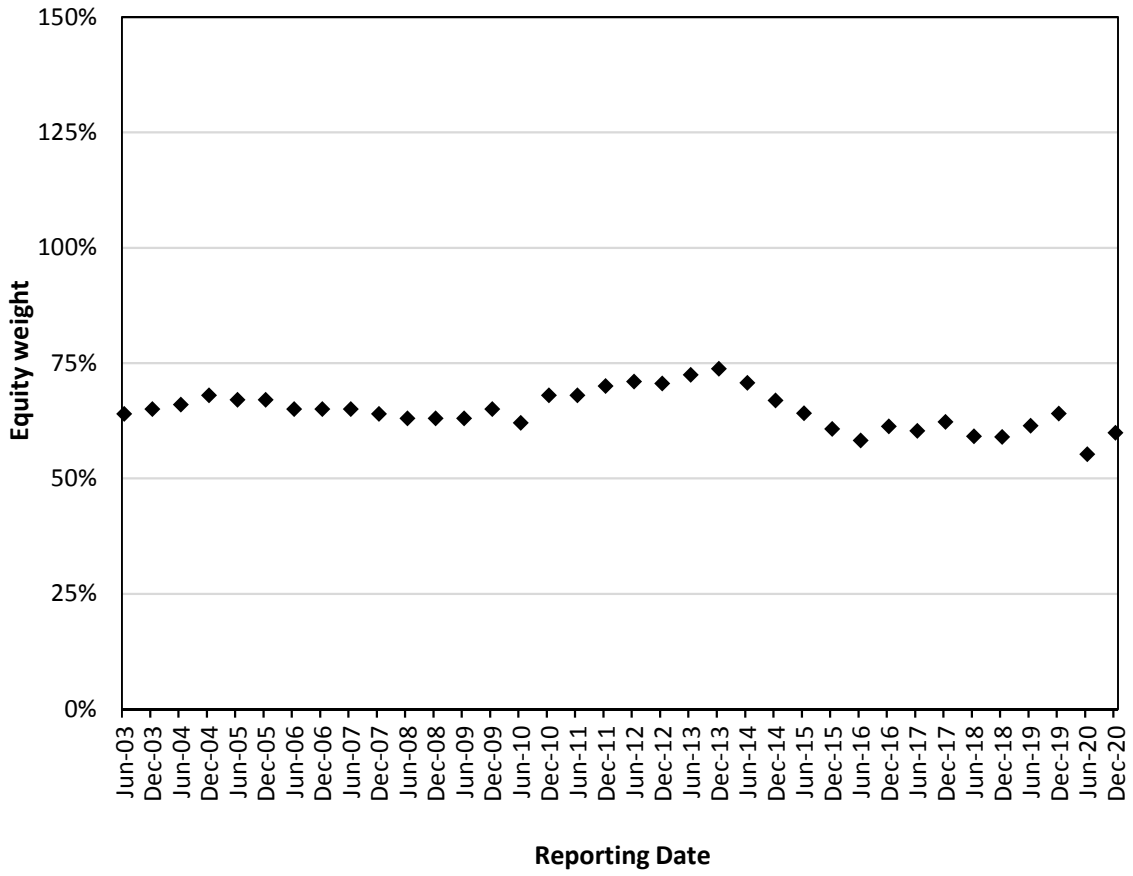


Figure B4: Equity weights of an exemplary mutual fund

For the time period from 2003 to 2020, for which the fund's reports are available on the Edgar system of the Securities and Exchange Commission (SEC), the graph shows the reported equity weights of the mutual fund American Balanced Fund. As of June 2018, American Balanced Fund was the largest US balanced fund.[16] The equity share for American Balanced Fund is collected from the fund's SEC filings N-CSR (annual) and N-CSR (semi-annual). During these years, American Balanced Fund did not report positions in equity futures or options.

Interpretation: The equity weight of a representative asset allocation fund fluctuates in a range that is considerably smaller than the 0% to 150% range that the predictability literature often chooses to make its benchmark weights realistic.



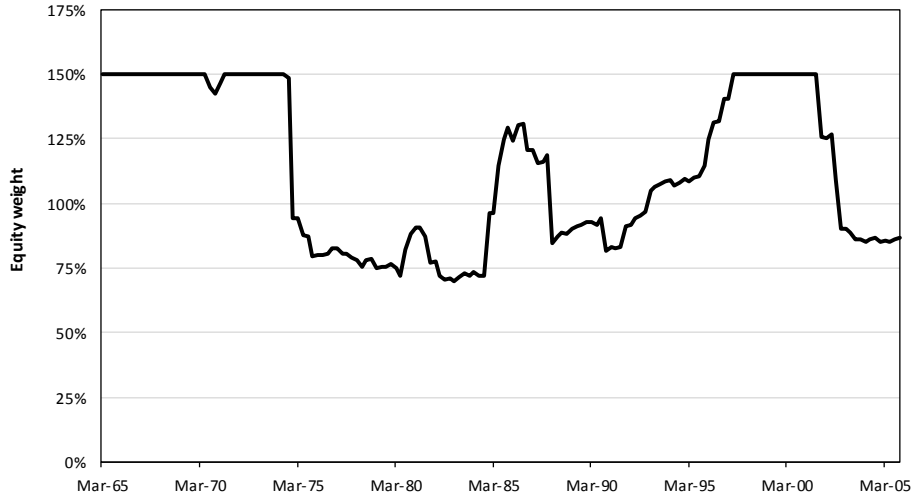
[16] Source: Pension&Investments (https://www.pionline.com/article/20181029/INTERACTIVE/181029977/the-largest-balanced-asset-allocation-mutual-funds-most-used-by-dc-plans)

Figure B5: Equity weights of the optimizing historical-mean benchmark in the out-of-sample periods of Rapach, Strauss, and Zhou (2010) and Rapach, Ringgenberg, and Zhou (2016).

Using the out-of-sample periods and assumptions made in the original papers, I determine the equity weight of the optimizing historical-mean benchmark. For Rapach, Strauss, and Zhou (2010), assumptions are: Trailing mean estimated since 1947:1; variance estimated using a rolling ten-year window; risk aversion $\gamma = 3$; weights limited to between 0 and 1.5. For Rapach, Ringgenberg, and Zhou (2010), assumptions are: Trailing mean estimated since 1973:12; variance estimated using a rolling ten-year window; risk aversion $\gamma = 3$; weights limited to between -0.5 and 1.5

Interpretation: The equity weights in historical-mean benchmarks of predictability studies can vary considerably.

Panel A: Equity weights of the historical-mean benchmark used in Rapach, Strauss, and Zhou (2010)



Panel B: Equity weights of the historical-mean benchmark used in Rapach, Ringgenberg, and Zhou (2016)

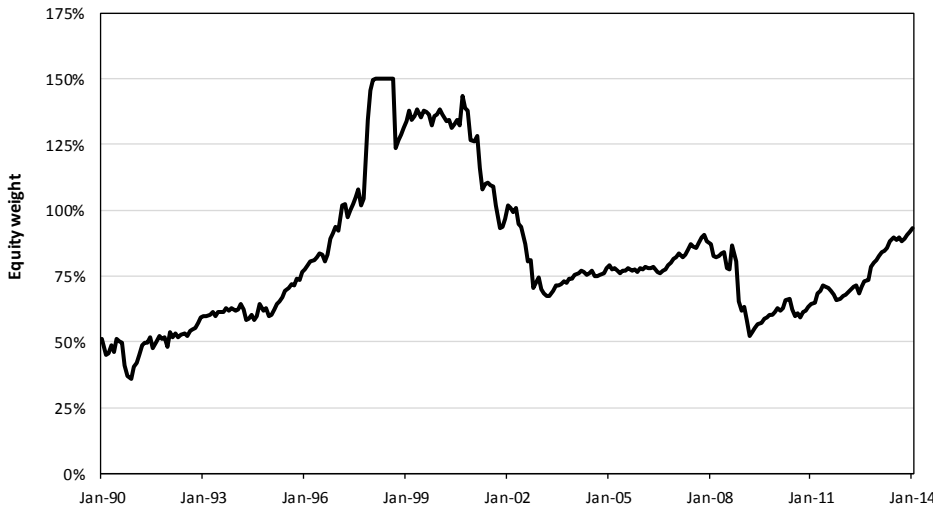
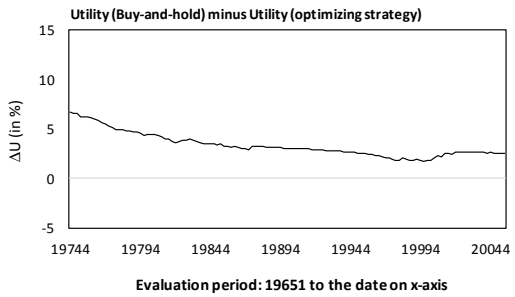


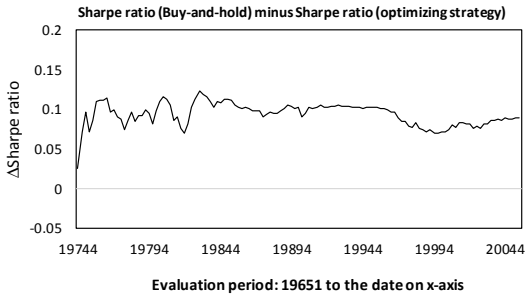
Figure B6: Performance of the buy-and-hold strategy relative to an optimizing strategy based on the historical mean—for expanding windows within the out-of-sample periods of the two replicated papers

The graphs show economic gains when switching from a strategy that optimizes the S&P 500 weight based on the historical mean to the buy-and-hold strategy. Utility gains Δu indicate the per annum gain of a mean-variance investor who switches from the optimizing strategy to the buy-and-hold strategy; Sharpe ratio differences are determined for the same switch. The analysis is conducted for expanding windows within the out-of-sample periods of two papers, using the return frequency and parameter assumptions made in these papers.

Interpretation: The superior performance of the buy-and-hold strategy is not only visible ex post; it would have also been visible to the imaginary investors in the out-of-sample periods.

Panel A: Economic gains from switching to the buy-and-hold strategy in the out-of-sample period of Rapach, Strauss, and Zhou (2010)





Panel B: Economic gains from switching to the buy-and-hold strategy in the out-of-sample period of Rapach, Ringgenberg, and Zhou (2016)