DEAKIN UNIVERSITY AUSTRALIA

Faculty of Business and Law

School of Accounting, Economics and Finance

School Working Paper

Financial Econometrics Series

SWP 2012/01

Does the Choice of Estimator Matter when Forecasting Returns?

J. Westerlund and P.K. Narayan

Does the Choice of Estimator Matter when Forecasting Returns?†

Joakim Westerlund[a,*], Paresh Narayan[a]

[a]*Deakin University, Faculty of Business and Law, School of Accounting, Economics and Finance, Melbourne Burwood Campus, 221 Burwood Highway, VIC 3125, Australia.*

This version: May 11, 2012

Abstract

While the literature concerned with the predictability of stock returns is huge, surprisingly little is known when it comes to role of the choice of estimator of the predictive regression. Ideally, the choice of estimator should be rooted in the salient features of the data. In case of predictive regressions of returns there are at least three such features; (i) returns are heteroskedastic, (ii) predictors are persistent, and (iii) regression errors are correlated with predictor innovations. In this paper we examine if the accounting of these features in the estimation process has any bearing on our ability to forecast future returns. The results suggest that it does.

JEL classification: C22; C23; G1; G12

Keywords: Predictive regression; Stock return predictability; Heteroskedasticity; Predictor endogeneity

1. Introduction

The use of financial ratios to predict returns has attracted much attention, and many studies have shown that ratios such as dividend–price, price–earnings, dividend–payout and book-to-market are able to predict future stock returns (see for example Campbell

†The first author would like to thank the Jan Wallander and Tom Hedelius Foundation for financial support under research grant number P2005–0117:1.

*Corresponding author. Tel.: +61 3 924 46973; fax: +61 3 924 46283.
E-mail addresses: j.westerlund@deakin.edu.au (Joakim Westerlund), paresh.narayan@deakin.edu.au (Paresh Narayan)

and Shiller, 1988a, 1988b, 1998; Fama and French, 1988; Kothari and Shanken, 1997; Lamont, 1998; Chen, 2009).[1] The prevailing tone in the literature is perhaps best summarized by Lettau and Ludvigson (2001, page 842): "It is now widely accepted that excess returns are predictable by variables such as dividend–price ratios, earnings–price ratios, dividend–earnings ratios, and an assortment of other financial indicators."

However, while there is some evidence of predictability, there is also evidence to the contrary, and in recent years this has become a major source of tension in the literature. In particular, while most evidence in favor of predictability are based on in-sample tests, the little out-of-sample evidence that exists is mainly negative. This disparity makes an overall assessment of return predictability difficult. The following passage, taken from Welch and Goyal (2008, page 1455), serves as an illustration: "The literature is difficult to absorb. Different articles use different techniques, variables, and time periods. Results from articles that were written years ago may change when more recent data is used. Some articles contradict the findings of others. Still, most readers are left with the impression that 'prediction works' – although it is unclear exactly what works."

In an attempt to sort out "what works", Welch and Goyal (2008) reconsider much of the empirical evidence reported in the literature. They find that most commonly used predictors are unable to deliver consistently superior out-of-sample forecasts of the US equity premium relative to a simple forecast based on the historical average. Similar conclusions are drawn by Bossaerts and Hillion (1999), who did not find any evidence of out-of-sample predictability in a collection of industrialized countries for a number of predictors. Hence, based on these results it would appear as that nothing works.

Amid this debate, in this paper we ask to what extent the out-of-sample forecasting performance is influenced by the choice of estimator of the predictive regression? This is a relevant question, because some of the differences in the literature may well be due to estimation problems. Indeed, as Rapach et al. (2010, page 288) point out:

> The lack of consistent out-of-sample evidence in Welch and Goyal indicates the need for improved forecasting methods to better establish the empirical reliability of equity premium predictability.

[1] There is a related branch of the literature which considers currency volatility predictability (see Scott and Tucker, 1989) and whether stock markets predict real activity (see Choi et al., 1999).

One of the potential problems is that many of the predictors are highly persistent, and their innovations tend to be correlated with return innovations. As Stambaugh (1999) shows, this combination of persistency and endogeneity can be quite devastating in that it induces a small-sample bias in the conventional ordinary least squares (OLS) estimator. Another potential problem is that returns are highly volatile. Indeed, one of the most well-documented features of financial time series is that returns are highly heteroskedastic. Thus, even if returns are predictable, to the extent that the heteroskedasticity is strong enough to dwarf the signal coming from the predictors, this might be difficult to detect.

Our approach to this issue is as follows. We consider monthly US time series data covering the period January 1871 to December 2008. The variables included are excess returns, the dividend–price ratio (DP), dividend yield (DY), the earnings–price ratio (EP), and the dividend–payout ratio (DE). Three estimators are applied to these data; (1) the OLS estimator, (2) the bias-adjusted OLS (AOLS) estimator of Lewellen (2004), and (3) the feasible generalized least squares (FGLS) estimator of Westerlund and Narayan (2011). The first estimator is included because it is the workhorse of the industry, the second is included because of its popularity, and the third is included because it is relatively new and therefore represents the state of the art. The difference between the estimators lies in their ability to accommodate possible heteroskedasticity, endogeneity and persistency of the regressor. While the OLS estimator is least general and does not account for any of these features, FGLS estimator is most general and accounts for all three features. The AOLS estimator lies somewhere in between and corrects for the endogeneity and persistency of the regressor.

The results suggest that the choice of estimator does make a difference, and that the FGLS estimator generally performs best. The results are shown to be significant not only from a statistical point of view, but also from an economic point of view. Moreover, while there is some variation coming from the choice of predictor and forecasting horizon, our results seem to be rather robust to the choice of sample period.

2. Econometric discussion

2.1. The predictive regression

As mentioned in Section 1, certain empirical features that characterize predictive regressions can make inference difficult and it is therefore important that these features are acknowledged already at the modeling stage. Our starting point is the following system of equations:

$$r_{t+h} = \alpha + \beta x_t + \epsilon_{t+h}, \tag{1}$$

$$x_{t+1} = \mu(1-\rho) + \rho x_t + \varepsilon_{t+1}, \tag{2}$$

where $|\rho| \leq 1$. This is the prototypical predictive regression model that has been widely used in the financial economics literature, in which x_t is a variable believed to be able to predict the h-period-ahead value of excess returns, r_{t+h}. In our case, x_t is either DP, DY, EP or DE. Thus, in this model testing the null hypothesis of no predictability is equivalent to testing the restriction that $\beta = 0$. As in previous studies, it is reasonable to assume that the correlation between ϵ_t and ε_t is negative. For example, if x_t is DY, then an increase in the stock price will lower dividends and raise returns. In order to capture endogenous effects of this sort, the following relationship between the error terms is assumed:

$$\epsilon_t = \gamma \varepsilon_t + \eta_t, \tag{3}$$

where ε_t and η_t are mean zero and uncorrelated with each other. The variances of these errors are typically assumed to be constant over time. However, this does not fit well with the fact that returns are almost always found to be heteroskedastic. The most popular approach by far to model this type of behavior is to assume an autoregressive conditional heteroskedasticity (ARCH) model, which in the case of η_t can be written as

$$\text{var}(\eta_t | I_{t-1}) = \sigma^2_{\eta t} = \lambda_0 + \sum_{j=1}^{q} \lambda_j \eta^2_{t-j}, \tag{4}$$

where I_t is the information available at time t. A similar equation is assumed to apply to $\text{var}(\varepsilon_t | I_{t-1}) = \sigma^2_{\varepsilon t}$. Provided that $\lambda_0 > 0$ and $\sum_{j=1}^{q} \lambda_j < 1$, this implies that the unconditional variance of η_t can be written in terms of the coefficients of (4) as $\text{var}(\eta_t) = \sigma^2_\eta = \frac{\lambda_0}{(1-\sum_{j=1}^{q} \lambda_j)}$ with $\text{var}(\varepsilon_t) = \sigma^2_\varepsilon$ having a similar representation. The unconditional variance of the composite error term ϵ_t can now be written as $\text{var}(\epsilon_t) = \sigma^2_\epsilon = \gamma^2 \sigma^2_\varepsilon + \sigma^2_\eta$.

Having laid out the model of interest, we now turn to the effects of endogeneity, persistency and ARCH:

- The correlation between ϵ_t and ε_t is given by $\rho_{\epsilon\varepsilon} = \gamma \frac{\sigma_\varepsilon}{\sigma_\epsilon}$ and is, as already mentioned, a source of major complication. The reason is that if $\rho_{\epsilon\varepsilon} \neq 0$, then x_t is no longer exogenous, thereby violating one of the most important OLS assumptions; if x_t is endogenous the OLS estimator of β is no longer unbiased.

- The main effect of the persistency of x_t is to aggravate the OLS bias caused by the endogeneity. In fact, as Stambaugh (1999) shows, the OLS bias is given by $-\gamma(1 + 3\rho)/T$, suggesting that while decreasing in T, the bias is increasing in γ and ρ. Moreover, the persistency is only a problem to the extent that x_t is indeed endogenous such that $\gamma \neq 0$.

- While unattended endogeneity and persistency are matters of bias, unattended ARCH is a matter of efficiency. Indeed, one of the most well-known results from classical regression theory is that OLS is inefficient in the presence of heteroskedasticity. In agreement with this, Westerlund and Narayan (2011) show that there are important efficiency gains to be made by accounting for the ARCH information.

2.2. The estimators

In view of the above discussion, the question of how to estimate the predictive regression in (1) naturally arises. One way is to simply ignore the issues of bias and inefficiency altogether, and run OLS. However, given the potential problems involved, this approach is clearly suboptimal, and Lewellen (2004) has therefore proposed a bias-adjusted estimator that deals with the first issue. The idea is to make (1) conditional on ε_t by substituting from (2) and (3), leading to the following augmented test regression:

$$r_{t+h} = \theta + \beta x_t + \gamma(x_{t+h} - \rho x_{t+h-1}) + \eta_{t+h}, \tag{5}$$

where $\theta = \alpha - \gamma\mu(1 - \rho)$. But while this removes the correlation between the regression error and the regressors, since ρ is unknown, (5) is not really feasible, and Lewellen (2004) therefore suggests replacing the true ρ with a "guess". Let ρ_0 denote this guess. The feasible version of (5) can be written as

$$r_{t+h} = \theta + \beta_0 x_t + \gamma(x_{t+h} - \rho_0 x_{t+h-1}) + \eta_{t+h}, \tag{6}$$

where $\beta_0 = \beta - \gamma(\rho - \rho_0)$ has the interpretation of a "bias-adjusted slope coefficient", which can be estimated by simply applying OLS to (6). The FGLS estimator is based on the same regression and is therefore very similar in spirit. One difference between the two estimators is the treatment of the persistency of x_t. In particular, while Lewellen (2004) assumes that $\rho_0 = \rho = 0.9999$, Westerlund and Narayan (2011) assume that $\rho = 1 + \frac{c}{T}$, where $c \leq 0$ is a drift parameter that measures the degree of persistency in x_t.[2] If $c = 0$, then x_t has an exact unit root, whereas if $c < 0$, then x_t is locally stationary in the sense that ρ approaches one from below as T increases. Thus, while Lewellen (2004) requires x_t to be stationary, which is quite restrictive, Westerlund and Narayan (2011) do not. Moreover, even if x_t happens to be stationary, the Lewellen (2004) assumption that ρ takes on the value 0.9999 is completely arbitrary.[3]

Another difference is the treatment of potential ARCH. As already mentioned, the AOLS estimator is based on running OLS on (6) with $\rho_0 = 0.9999$, which means that any information contained in the ARCH structure of the errors is effectively ignored. The FGLS estimator, on the other hand, weights all the data by $\frac{1}{\sigma_{\eta t}}$ and is therefore able to exploit this information. However, it is still the same equation that is estimated, although in this case with $\rho_0 = 1$. Note, though, that in this case the fact that ρ_0 is set to 1 does not mean that ρ must be 1 too; it just means that under the above "local-to-unity" assumption ρ is so close to 1 that we can just as well set $\rho_0 = 1$.

2.3. *Out-of-sample forecasting*

As alluded in Section 1, most existing evidence of the predictability of returns is based on in-sample results and there are only a hand-full of studies that consider out-of-sample results. The idea of out-of-sample forecasting is to save the last P observations for the forecasting evaluation and to fit the predictive regression to the first $T_0 = T - P$ observations only. One then generates P forecasts with recursively updated model estimates.

To illustrate the effect of the choice of estimator, assume for simplicity that $\alpha = 0$

[2]For similar assumptions, see Campbell and Yogo (2006) and the references provided therein.

[3]The price of the greater flexibility of the local-to-unity specification of ρ is that the theoretical analysis becomes more involved. In particular, it introduces c as an additional nuisance parameter in the asymptotic distribution of the estimated predictive slope.

and denote by $\hat{\beta}_{AOLS}$ and $\hat{\beta}_{FGLS}$ the AOLS and FGLS estimators of β, respectively. The associated predictive values are denoted $\hat{r}_{t+h}^{AOLS} = \hat{\beta}_{AOLS} x_t$ and $\hat{r}_{t+h}^{FGLS} = \hat{\beta}_{FGLS} x_t$, where $t = T_0 + 1, ..., T$. In this notation the OLS-based forecast error can be written as

$$\hat{r}_{t+h}^{AOLS} - r_{t+h} = (\hat{\beta}_{AOLS} - \beta) x_t + \epsilon_{t+h}$$

with a similar representation applying to $\hat{r}_{t+h}^{FGLS} - r_{t+h}$. The ratio of the mean squared forecast errors (MSFE) therefore becomes

$$\frac{E[(\hat{r}_{t+h}^{FGLS} - r_{t+h})^2]}{E[(\hat{r}_{t+h}^{AOLS} - r_{t+h})^2]} = \frac{E[(\hat{\beta}_{FGLS} - \beta)^2 x_t^2] + \sigma_\epsilon^2}{E[(\hat{\beta}_{AOLS} - \beta)^2 x_t^2] + \sigma_\epsilon^2} + o_p(1),$$

where the $o_p(1)$ term is negligible as T tends to infinity. In the presence of ARCH the FGLS estimator is more efficient than the AOLS estimator, suggesting that $E[(\hat{\beta}_{FGLS} - \beta)^2 x_t^2]$ should in general be smaller than $E[(\hat{\beta}_{AOLS} - \beta)^2 x_t^2]$. In order to appreciate this, note that

$$\begin{aligned} E[(\hat{\beta}_{FGLS} - \beta)^2] &= E[(\hat{\beta}_{FGLS} - E(\hat{\beta}_{FGLS}) + (E(\hat{\beta}_{FGLS}) - \beta))^2] \\ &= E[(\hat{\beta}_{FGLS} - E(\hat{\beta}_{FGLS}))^2] + (E(\hat{\beta}_{FGLS}) - \beta)^2, \end{aligned}$$

where the first term on the right-hand side is simply the variance of the FGLS estimator, while the second is the squared bias. A similar representation holds for the AOLS estimator. Hence, the relative MSFE depends on both the efficiency and bias of the two estimators. The FGLS is more efficient, which means that the MSFE ratio should be smaller than one. However, because of the consistency of the two estimators, $(\hat{\beta}_{FGLS} - \beta)$ and $(\hat{\beta}_{AOLS} - \beta)$ will tend to zero as T grows, which means that the MSFE ratio is asymptotically one. Thus, while the FGLS is expected to perform better in small samples, in large samples this is not the case.

It should be noted here that while the OLS estimator has been used before in out-of-sample studies of return predictability, to the best of our knowledge, this is the first time the AOLS and FGLS estimators has been used for this purpose. In particular, while the work of Lewellen (2004) has attracted much attention, what researchers tend to use is his predictability test, and not the associated slope estimator. Westerlund and Narayan (2011) also focus on the testing problem. But, again, there is nothing that prevents us from using their proposal for forecasting purposes. Thus, while first considered by

Lewellen (2004), and Westerlund and Narayan (2011), in these papers the AOLS and FGLS estimators are used mainly for the purpose of constructing t-statistics for the null hypothesis of no predictability, and not for point estimation and/or forecasting.

2.4. Simulation evidence

A small-scale simulation study was conducted to assess the relative performance of the three estimators considered. The conditional mean equation is generated from (1)–(4) with $\alpha = h = 1$, $\mu = 0$ and $\eta_t \sim N(0, \sigma^2_{\eta t})$ and $\varepsilon_t \sim N(0, \sigma^2_{\varepsilon t})$. The ARCH equations driving $\sigma^2_{\eta t}$ and $\sigma^2_{\varepsilon t}$ are assumed to be of the first order with equal intercept and slope coefficients, henceforth denoted λ_0 and λ_1, respectively. Also, in order to ensure that $\sigma^2_\eta = \sigma^2_\varepsilon = 1$ are kept fixed, we set $\lambda_0 = 1 - \lambda_1$. The value of β did not affect the results and we therefore set $\beta = 1$. Also, as for ρ, we follow Westerlund and Narayan (2011), and set $\rho = 1 + \frac{c}{T}$, where c can be either zero, such that x_t is unit root non-stationary, or -2, such that x_t is locally stationary.

By setting λ_1 and γ, we can control the degree of ARCH and endogeneity. The values considered for γ are zero and -2, corresponding to the cases of no and strong endogeneity. Note in particular that $\gamma = -2$ implies $\rho_{\epsilon\varepsilon} \simeq -0.9$, which is consistent with the largest correlations in our data, and also with the parametrization of Lewellen (2004, Figure 1). The values considered for λ_1 are zero and 0.6, corresponding to the cases of no and strong ARCH effects. In fact, since under normality the condition for the standardized forth moments of η_t and ε_t to be finite is given by $\alpha_1 < 1/\sqrt{3} \simeq 0.6$, in the high ARCH effect case the kurtosis is borderline infinite. Since the asymptotic theory of the estimators considered here assumes the existence of at least fourth moments, $\lambda_1 = 0.6$ provides a limit for how much ARCH we can reasonably allow for without risking non-sensical results. The number of replications is set to $3,000$.

The results from the average of the simulated root-MSFE ratios for the FGLS estimator versus the OLS and AOLS estimators are reported in Table 1. The findings are largely as expected. First, all three estimators seem to perform very similarly when $\lambda_1 = \gamma = 0$. Secondly, when $\gamma = -2$ we see that while the FGLS and AOLS estimators tend to perform similarly, the OLS estimator is strictly dominated by the FGLS estimator, suggesting that the accounting for endogeneity can be important in practice. Third, when $\lambda_1 = 0.6$ the FGLS estimator generally outperforms the competition, meaning that the accounting for ARCH is also important. The extent of the performance gain is decreasing in T, which is in agreement with the discussion of Section 2.3.

3. Empirical results

3.1. Data

The data used are taken from Robert Shiller's webpage and are for the period January 1871 to December 2008. We download data on real returns, real earnings and real dividends on the S&P 500 index. Excess returns is computed by subtracting the three month Treasury bill rate from returns. The predictors are defined as follows:

- DP is defined as the difference between log dividends and the log of stock prices, where dividends are measured using a one-year moving sum;
- DY is the difference between log dividends and the log of the one-period lagged stock price;
- EP is the difference between log earnings and the log of stock prices, where earnings are measured using a one-year moving sum;
- DE is the difference between log dividends and log earnings.

From the full sample we consider three out-of-sample forecasting periods; (1) January 1965–December 2008, (2) January 1976–December 2008, and (3) January 2001–December 2008. Our choice of considering subperiods is motivated by the previous literature arguing that out-of-sample predictability may be driven in part by the choice of the sample period (see for example Rozeff, 1984; Fama and French, 1988; Lettau and Nieuwerburgh, 2008; Boudoukh et al., 2008). Because this choice is arbitrary, it leaves open the question of data mining. The post-1976 period is motivated by Welch and Goyal's (2008) observation that following the oil price shock of the mid-1970's, the predictability of many economic variables weakened. The motivation for the post-2001 period is to gauge whether return predictability has been affected by the "technology bubble" of the last decade (Rapach and Wohar, 2006).

3.2. Preliminary results

When we examine some selected descriptive statistics of the data, we make a number of observations suggesting that the three out-of-sample forecasting periods considered feature different behavior. In Table 2, we report the mean, coefficient of variation, skewness, kurtosis, and autocorrelations based on the squared variables. These statistics are reported for the full 1871–2008 period as well as for the three out-of-sample periods. Consider returns. Over the 1965–2008 period returns were negative, they became positive in the 1976–2008 period, and in the most recent 2001–2008 period average returns were again negative. We also see that returns were lowest in the most recent period. Also, the volatility of returns has declined substantially over the years. A similar pattern emerges for the four predictors; the variables have on average become more negative over the years and their volatility has declined.

In the last four columns of Table 2, we report the autocorrelations associated with the square of each of the variables, which can be thought of as estimated ARCH coefficients. We notice that while for the predictors the autocorrelations are significant, for returns, except when considering the full 1871–2008 period, there is a clear tendency for the p-values to increase with more distant lags. The evidence of ARCH is therefore stronger for the predictors than for returns. In any event, the implication here is that both the predictors and returns are characterized by ARCH.

In Table 3 we report some augmented Dickey–Fuller (ADF) test results. The test is implemented both with and without a linear time trend; however, since the results were very similar we only report the results for the model with a constant but no trend. The order of the lag augmentation is chosen by the Schwarz information criterion with a maximum of eight lags. The results reveal that while for the full sample there is strong evidence against the unit root null hypothesis for all four predictors, for the three subperiods there is no such evidence. By comparison, the unit root null is consistently rejected for returns. In other words, while returns can be considered as stationary, the predictors are highly persistent and borderline unit root non-stationary, suggesting that our preference of modeling x_t as a near unit root variable seems quite reasonable.

In order to also get a feeling for the extent of the endogeneity we computed the correlation between the OLS residuals of the predictive regression in (1) and the first difference of the predictors. DP leads to the largest correlation of -0.95, while DE leads to the smallest correlation of -0.12. DY and EP lie in between with correlations of -0.29 and -0.80, respectively. Since all correlations are also significant, even at the 1% level, we take this as strong evidence of endogeneity.

The evidence reported so far suggests that (1) the predictors are highly persistent, (2) the residuals of the predictive regression are correlated with the first-differenced predictors, and (3) the data are heteroskedastic. This means that OLS is not only inefficient but also biased, making it a poor candidate not only for inference but also for forecasting (see for example Lanne, 2002). We are therefore tempted to disregard the OLS estimator already at the outset. For completeness, however, in Sections 3.3 and 3.4 the OLS results are included, although readers are advised to interpret them with care. The AOLS estimator is unbiased, which makes it a better choice than OLS; however, it is still inefficient. The FGLS estimator is the only one that is expected to be both unbiased and efficient, and is expected to lead to superior forecasting performance when compared to its two competitors.

3.3. Out-of-sample results

The evaluation of the out-of-sample forecasting results is done on a pair-wise basis using the well-known Theil U and Diebold–Mariano statistics, and for four horizons; $h = 1$, $h = 3$, $h = 6$ and $h = 12$. The forecasting is carried out as explained in Section 2.3, by recursively expanding the sample. The OLS and AOLS estimators are implemented as described in Section 2.2. As for the FGLS estimator, the weight $\frac{1}{\sigma_{\eta t}}$ needs to be replaced with an estimate. The approach used here is the same as the one in Westerlund and Narayan (2011), and proceeds as follows. We begin by applying OLS to (6) with $\rho_0 = 1$. This gives us the OLS estimator $\hat{\eta}_t$ of η_t. The estimator of $\sigma^2_{\eta t}$ is then obtained as the fitted value from a regression of $\hat{\eta}^2_t$ onto a constant and q of its own lags, where q is selected using a recursive general-to-specific t-test on the last lag with a significance level of 5%.[4]

[4]The idea here is to simply rewrite (4) as $\eta^2_t = \lambda_0 + \sum_{j=1}^{q} \lambda_j \eta^2_{t-j} + (\eta^2_t - \sigma^2_{\eta t})$, where $(\eta^2_t - \sigma^2_{\eta t})$ is a martingale difference sequence. This suggests that a natural approach to obtain $\hat{\sigma}^2_{\eta t}$ is to first obtain $\hat{\eta}_t$ and then to retrieve $\hat{\sigma}^2_{\eta t}$ as the fitted value from a regression of $\hat{\eta}^2_t$ onto $(1, \hat{\eta}^2_{t-1}, ..., \hat{\eta}^2_{t-q})$. This procedure is explained in Westerlund and Narayan (2011).

We begin by considering the results reported in Table 4 for the comparison between the AOLS and OLS estimators. If the Theil U and Diebold–Mariano statistics are close to one and zero, respectively, then the two estimators are equally good at predicting returns, whereas if the statistics are larger than these values, OLS is best. When we use DP as a predictor, we see that there is not much evidence against the null hypothesis of equal predictability. The Theil U and Diebold–Mariano statistics are generally larger than one and zero, respectively, suggesting that OLS performs best; however, the difference is not large enough for a rejection by the Diebold–Mariano test. This means that both estimators have equal predictive power. This is true for all horizons. The conclusion for DY is the same, although in this case the evidence leans more towards the AOLS estimator with the Theil U and Diebold–Mariano statistics being smaller than one and zero, respectively.

While still weak the evidence against the equal predictability null is stronger for EP and DE. For EP we count one rejection at the 10% level, at the $h = 1$ horizon in the most recent 2001–2008 out-of-sample period. Since both statistics are larger than their expected values under the null, OLS is preferred. The relatively good performance of the OLS estimator is confirmed by the DE results, but now we count three rejections at the same level of significance.

Consider next the results reported in Table 5 for the comparison between the FGLS and OLS estimators. Now there are as many as seven rejections. However, the evidence is rather mixed and there is no clear-cut winner. For example, in case of DP the Diebold–Mariano statistic results in two rejections, one favors OLS, while the other favors the FGLS estimator. This picture is reinforced if we look at the magnitude of the Theil U and Diebold–Mariano statistics; there are about as many instances when they are larger than one and zero as when they are less than one and zero. By contrast, the results for DY and EP are more supportive of the FGLS estimator, especially when $h = 12$. The overall impression so far is therefore that the FGLS estimator performs best. However, the results for DE are more favorable towards OLS with two rejections in its favor.

Our last comparison is between the FGLS and AOLS estimators. The results are reported in Table 6. The first thing to notice is that the Theil U and Diebold–Mariano statistics are generally lower than one and zero, respectively, which in this case means evidence in favor of the FGLS estimator. Overall, we count seven rejections at the 10% level, of which six provide evidence in support of FGLS. Three of these are for DP, one is for DY and the remaining two are for DE. We also see that the evidence in favor of the FGLS estimator is stronger at shorter horizons than at longer horizons.

It should be pointed out that while the OLS estimator actually turns out to perform rather well here, as noted in Section 3.2, it is inefficient and biased, which do not directly inspire confidence in a potential investor. We will therefore not consider it any further.

3.4. Economic significance

The goal of this section is to analyze any economic significance available to investors from the forecasting performance of the FGLS, AOLS and OLS estimators. Our approach is similar to the one of Rapach et al. (2010), Campbell and Thompson (2008), and Marquering and Verbeek (2004). More specifically, we first compute the average utility for a mean-variance investor with relative risk aversion parameter ϕ who allocates her portfolio between stocks and risk-free bills using h-period-ahead forecasts of returns based on the FGLS estimator. The utility function is assumed to be given by $E(r_{t+h}|I_t) - \frac{\phi}{2}\text{var}(r_{t+h}|I_t)$, which can be mimicked out-of-sample using the estimated mean and variance of the FGLS-based return forecasts. We then compute the average utility for the same investor when using the AOLS and OLS estimators to obtain the forecasts.

We compute annual average utilities based on two risk aversion parameters, $\phi = 3$ and $\phi = 6$, and two horizons, $h = 1$ and $h = 12$. The utility gain is estimated as the difference in utility between the FGLS and AOLS or OLS forecasts, and we multiply this difference by 1,200 to express it in average annualized percentage return. The utility gain can be interpreted as the portfolio management fee that an investor would be willing to pay to have access to the former estimator.

The results are reported in Table 7. We begin by considering the results from the comparison with the AOLS estimator. Earlier we showed that the FGLS estimator results in more precise forecasts than the AOLS estimator. If this is true use of the FGLS estimator should lead to positive utility gains, and this is exactly what we see in the table. We also see that the utility gains associated with the FGLS forecasts can be quite sizable, especially when ED is used as a predictor, in which case some the utility gains are well above 1%. The only exception is for DE in the 1965–2008 period, in which case the investor would be better off by using the AOLS estimator. However, given the smallness of the utility gain in this case, one could probably just as well use the FGLS estimator.

Consider next the utility gains from using FGLS rather than OLS. Two observations are worth highlighting. First, most utility gains are positive, favoring the FGLS estimator over the OLS estimator. Moreover, these results are robust to different out-of-sample periods, horizons, and risk aversion parameters. Second, the only predictor for which, at least to some extent, the OLS is relatively favorable is DY. We notice that only when the risk aversion parameter is set to three, over both the short and long horizons utility gains are negative suggesting that the OLS estimator is relatively better. However, we note that this result is reversed in favor of the FGLS when the risk aversion parameter is set to six, suggesting that the results in favor of OLS with the DY predictor are not robust.

3.5. Robustness checks

In this section we check the robustness of our results in two regards. First, we examine whether the superior performance of the FGLS estimator remains also when compared to the historical average. Second, we examine if the results are robust to the choice of lag length.

Many studies of predictive ability take as their benchmark the historical average forecast (see for example Rapach et al., 2005; Flood and Rose, 2010). As a first robustness check we therefore compare this forecast with that of our FGLS estimator, which we have demonstrated to perform best among the estimators considered here. One problem towards this end is that because the models are now nested the conventional Diebold–Mariano statistic does not have its usual asymptotic standard normal distribution (see Clark and McCracken, 2001; McCracken, 2007). In order to account for this fact we follow Rapach and Weber (2004), and Rapach *et al*. (2005), among others, and use the MSE–*F* and ENC–NEW statistics of McCracken (2004), and Clark and McCracken (2001), respectively. Moreover, following the recommendation of Clark and McCracken

(2004) the p-values are obtained by using the bootstrap.[5]

The results are reported in Table 8. The first thing to notice is that most of the test statistics are positive, suggesting that the FGLS estimator generally outperforms the historical average. The evidence is particularly strong for DP and DY, in which case there is significant evidence in favor of FGLS at all horizons considered. The evidence for EP and DE is more mixed with a majority of p-values being greater than 0.1, suggesting that neither model is superior.

Consider next the effect of the selection of the number of lags to use in the estimation of the ARCH equation. In the second half of Table 3, we report some ARCH test results. Essentially, in order to remove the effect of serial correlation, we begin by running an autoregression with 12 lags. Following this, we undertaken a residual-based Lagrange multiplier test for heteroskedasticity. The null hypothesis is that there is no ARCH effects. Two lag lengths are considered, six and 12. According to the results, the null of no ARCH must be rejected for both lag lengths and for all variables. The only exception is for returns in the most recent 2001–2008 forecasting period, where the null cannot be rejected at the 10% level.

The main implication here is that an ARCH model might not be enough to capture the heteroskedasticity. As a response to this, Westerlund and Narayan (2011) consider a generalized ARCH (GARCH) model, which they show can be approximated by means of a high-order ARCH model. As a robustness check, in this section we therefore consider an ARCH model with 12 lags and compare the predictive performance of the resulting FGLS estimator with that of the AOLS estimator.

The results for the comparison between the GARCH–FGLS and AOLS estimators are not reported but we briefly describe them. As in Table 6, we find that the Theil U and Diebold–Mariano statistics are generally lower than one and zero, respectively, which we take as evidence in favor of the FGLS estimator. In fact, the results are very similar to those reported in Table 6. The only main exception is DE, for which the evidence in support of the AOLS estimator is now much stronger than before with four rejections at the positive side.

As a final robustness test, at a suggestion of a referee of this journal, we checked the

[5] A detailed description is available upon request from the corresponding author.

relative out-of-sample forecasting performance of the FGLS estimator when fixing ρ_0 in the same way as in AOLS. The results for the case when $h = 1$ are reported in Table 9. Consistent with the results based on assuming that is local-to-unity, we see that the Theil and Diebold–Mariano statistics tend to favor the FGLS estimator when using DP, DY, and EP as predictors. This is true regardless of the out-of-sample forecasting period considered.

4. Concluding remarks

This paper studies the relative forecasting performance of three estimators, namely the OLS, AOLS and FGLS estimators. While OLS is least general in the sense that it ignores the problems of heteroskedasticity, and predictor endogeneity and persistency, the FGLS estimator is most general and accounts for all three problems. The AOLS estimator accounts for the endogeneity and persistency of the predictor, and therefore lies somewhere in between. The first problem is a matter of bias, while the second is a matter of efficiency, but both are expected to be important for forecasting purposes. Of course, in-sample the OLS estimator is expected to perform best notwithstanding its potential bias and lack of efficiency. The relative out-of-sample performance of these estimators is, however, still an open issue.

The data we use include observations on excess returns on the S&P 500 index and four common predictors, ED, DY, EP and DE. The sample starts in January 1871 and ends in December 2008, which means that we have no less than 1,655 monthly observations to our disposal. The results from the pair-wise out-of-sample forecasting evaluation suggest that the FGLS estimator generally performs best. The results from the comparison between the FGLS and AOLS estimators are particularly clear with the former estimator having the best performance by far. The results from the OLS comparison are more mixed, although overall the evidence tend to lean towards the FGLS estimator, especially when considering a relatively long forecasting horizon. This conclusion is robust to choice of out-of-sample period.

In an attempt to quantify the economic significance of the results from the forecasting comparison, we examine whether a mean-variance investor would be better off in terms of utility by using the FGLS estimator rather than the AOLS estimator. The results suggest that she will.

References

Bossaerts, P., Hillion, P., 1999. Implementing statistical criteria to select return forecasting models: What do we learn? Review of Financial Studies 12, 405–428.

Boudoukh, J., Richardson, M., Whitelaw, R. F., 2008. The myth of long-horizon predictability. Review of Financial Studies 21, 1577–1605.

Campbell, J. Y., Shiller, R. J., 1988a. The dividend-price ratio and expectations of future dividends and discount factors. Review of Financial Studies 1, 195–228.

Campbell, J. Y., Shiller, R. J., 1988b. Stock prices, earnings, and expected dividends. Journal of Finance 43, 661–676.

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

Campbell, J. Y., Thompson, S. B., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average. Review of Financial Studies 21, 1509–1531.

Campbell, J. Y., Yogo, M., 2006. Efficient tests of stock return predictability. Journal of Financial Economics 81, 27–60.

Chen, S.-S., 2009. Predicting the bear stock market: Macroeconomic variables as leading indicators. Journal of Banking and Finance 33, 211–223.

Choi, J. J., Hauser, S., Kopecky, K. J., 1999. Does the stock market predict real activity? Time series evidence from the G-7 countries. Journal of Banking and Finance 23, 1771–1792.

Clark, T. E., McCracken, M. W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110.

Clark, T. E., McCracken, M. W., 2004. Evaluating long-horizon forecasts. Unpublished manuscript.

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

Kothari, S. P., Shanken, J., 1997. Book-to-market, dividend yield, and expected market returns: A time series analysis. Journal of Financial Economics 44, 169–203.

Lamont, O., 1998. Earnings and expected returns. Journal of Finance 53, 1563–1587.

Lanne, M., 2002. Testing the predictability of stock returns. Review of Economics and Statistics 84, 407–415.

Lettau, M., Nieuwerburgh, S. V., 2008. Reconciling the return predictability evidence. Review of Financial Studies 21, 1607–1652.

Lettau, M., Ludvigson, S., 2001. Consumption, aggregate wealth, and expected stock returns. Journal of Finance 56, 815–849.

Lewellen, J., 2004. Predicting returns with financial ratios. Journal of Financial Economics 74, 209–235.

McCracken, M. W., 2004. Asymptotics for out-of-sample tests of causality. Working paper, University of Missouri-Columbia.

McCracken M.W.,West, K. D., 2002. Inference about predictive ability. In: Clements M. P., Hendry, D. F. (Eds.), A companion to economic forecasting. Blackwell, Oxford.

McCracken, M. W., 2007. Asymptotics for out-of-sample tests of Granger causality. Journal of Econometrics 140, 719–752.

Rapach, D. E., Weber, C. E., 2004. Financial variables and the simulated out-of-sample forecastability of U.S. output growth since 1985: An encompassing approach. Economic Inquiry 42, 717–738.

Rapach, D. E., Wohar, M. E., Rangvid, J., 2005. Macro variables and international stock return predictability. International Journal of Forecasting 21, 137–166.

Rapach, D. E., Wohar, M. E., 2006. In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. Journal of Empirical Finance 13, 231–247.

Rapach, D. E., Strauss, J. K., Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Review of Financial Studies 23, 821–862.

Rozeff, M. S., 1984. Dividend yields are equity risk premiums. Journal of Portfolio Management 11, 68–75.

Scott, E., Tucker, A. L., 1989. Predicting currency return volatility. Journal of Banking and Finance 13, 839–851.

Stambaugh, R. F., 1999. Predictive Regressions. Journal of Financial Economics 54, 375–421.

Mankiw, N. G., Shapiro, M. D., 1986. Do we reject too often? Economic Letters 20, 139–145.

Marquering, W., Verbeek, M., 2004. The economic value of predicting stock index returns and volatility. Journal of Financial and Quantitative Analysis 39, 407–429.

Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455–1508.

Westerlund, J., Narayan, P., 2011. Testing for predictability in heteroskedastic stock returns. Unpublished manuscript.

Table 1: Simulated root-MSFE ratios.

			$c = 0$		$c = -2$	
λ_1	γ	T	OLS	AOLS	OLS	AOLS
0	0	100	1.008	1.005	1.002	1.001
0	0	200	1.001	1.001	1.000	1.000
0	0	400	1.001	1.000	1.000	1.000
0	−2	100	0.902	1.003	0.978	1.001
0	−2	200	0.962	1.001	0.992	1.000
0	−2	400	0.987	1.000	1.000	1.000
0.6	0	100	1.001	1.000	1.003	1.003
0.6	0	200	0.999	0.999	0.999	0.998
0.6	0	400	0.999	0.999	1.000	1.000
0.6	−2	100	0.888	0.997	0.978	1.001
0.6	−2	200	0.957	0.998	0.988	0.999
0.6	−2	400	0.986	0.999	0.998	1.000

Notes: The numbers in the table are the average of the simulated root-MSFE ratios for the FGLS estimator versus the OLS and AOLS estimators. λ_1 and γ determine the extent of ARCH and endogeneity, respectively. c is the drift in the local-to-unity specification of x_t.

Table 2: Descriptive statistics.

Variable	Mean	CV	Skewness	Kurtosis	AR(1)	AR(6)	AR(24)
				1871–2008			
Returns	0.00	1646.80	−0.37	14.38	0.13 (0.00)	0.11 (0.00)	0.02 (0.00)
DP	−3.17	0.13	−0.89	4.00	1.00 (0.00)	0.96 (0.00)	0.81 (0.00)
DY	−3.17	0.13	−0.92	4.02	1.00 (0.00)	0.96 (0.00)	0.81 (0.00)
EP	−2.65	0.13	−0.09	3.34	0.98 (0.00)	0.89 (0.00)	0.61 (0.00)
DE	−0.53	0.53	0.30	3.57	1.00 (0.00)	0.92 (0.00)	0.53 (0.00)
				1965–2008			
Returns	−0.03	118.58	−1.02	6.93	0.15 (0.00)	0.04 (0.00)	−0.04 (0.53)
DP	−3.55	0.12	−0.45	2.30	1.00 (0.00)	0.97 (0.00)	0.85 (0.00)
DY	−3.55	0.12	−0.45	2.32	1.00 (0.00)	0.97 (0.00)	0.85 (0.00)
EP	−2.78	0.15	0.03	2.67	0.98 (0.00)	0.91 (0.00)	0.68 (0.00)
DE	−0.77	0.29	0.56	5.86	0.99 (0.00)	0.86 (0.00)	0.22 (0.00)
				1976–2008			
Returns	0.10	37.13	−1.14	8.08	0.16 (0.00)	0.04 (0.00)	−0.03 (0.85)
DP	−3.60	0.13	−0.17	1.82	1.00 (0.00)	0.97 (0.00)	0.86 (0.00)
DY	−2.40	0.13	−0.18	1.82	1.00 (0.00)	0.97 (0.00)	0.86 (0.00)
EP	−2.80	0.16	0.06	2.26	0.98 (0.00)	0.91 (0.00)	0.67 (0.00)
DE	−0.80	0.29	0.91	6.10	0.99 (0.00)	0.84 (0.00)	0.14 (0.00)
				2001–2008			
Returns	−0.70	6.01	−1.79	9.34	0.12 (0.20)	0.04 (0.90)	−0.02 (1.00)
DP	−4.10	0.05	0.32	4.40	0.92 (0.00)	0.62 (0.00)	0.10 (0.00)
DY	−4.11	0.05	0.03	3.92	0.92 (0.00)	0.65 (0.00)	0.12 (0.00)
EP	−3.20	0.09	−0.66	2.79	0.91 (0.00)	0.74 (0.00)	0.09 (0.00)
DE	−0.91	0.35	1.72	7.49	0.98 (0.00)	0.69 (0.00)	−0.30 (0.00)

Notes: CV refers to the coefficient of variation and AR(p) refers to the p^{th} autocorrelation in the squared variable. The values in parentheses are the p-values for the zero autocorrelation restriction.

Table 3: ADF and ARCH test results.

Variable	ADF	p-value	ARCH(6)	p-value	ARCH(12)	p-value
			1871–2008			
Returns	−26.16	0.00	32.35	0.00	24.66	0.00
DP	−2.71	0.07	1054.00	0.00	580.00	0.00
DY	−2.90	0.05	897.00	0.00	490.00	0.00
EP	−2.80	0.06	1129.00	0.00	573.00	0.00
DE	−3.25	0.02	3342.00	0.00	1673.00	0.00
			1965–2008			
Returns	−17.60	0.00	2.52	0.02	1.62	0.08
DP	−1.29	0.64	318.60	0.00	159.60	0.00
DY	−1.28	0.64	264.30	0.00	132.40	0.00
EP	−0.36	0.91	255.70	0.00	127.10	0.00
DE	−0.19	0.94	4308.00	0.00	2559.00	0.00
			1976–2008			
Returns	−15.14	0.00	2.51	0.02	1.63	0.08
DP	−1.06	0.73	205.70	0.00	99.79	0.00
DY	−1.06	0.73	152.40	0.00	73.49	0.00
EP	−0.12	0.95	161.90	0.00	79.51	0.00
DE	0.00	0.96	3271.00	0.00	1698.00	0.00
			2001–2008			
Returns	−8.05	0.00	0.70	0.65	0.68	0.76
DP	0.91	1.00	79.59	0.00	36.04	0.00
DY	0.97	1.00	46.37	0.00	22.06	0.00
EP	−0.17	0.94	32.23	0.00	12.92	0.00
DE	0.88	1.00	1036.00	0.00	535.30	0.00

Notes: ADF refers to the augmented Dickey–Fuller test, and ARCH(q) refers to a Lagrange multiplier test of the zero slope restriction is an ARCH regression of order q.

Table 4: Out-of-sample results for AOLS versus OLS.

		$h = 1$		$h = 3$		$h = 6$		$h = 12$	
Predictor	Sample	DM	Theil U	DM	Theil U	DM	Theil U	DM	Theil U
DP	1965–2008	2.436 (0.015)	1.011	1.490 (0.160)	1.013	1.494 (0.135)	1.013	0.008 (0.994)	1
	1976–2008	0.871 (0.384)	1.005	0.284 (0.777)	1.003	0.274 (0.784)	0.003	−0.759 (0.448)	0.989
	2001–2008	2.217 (0.033)	1.017	1.302 (0.193)	1.018	1.269 (0.205)	1.021	0.543 (0.587)	1.015
DY	1965–2008	−0.083 (0.934)	1	0.413 (0.679)	1.001	−0.184 (0.854)	0.999	−1.102 (0.279)	0.994
	1976–2008	−1.136 (0.256)	0.998	−0.570 (0.569)	0.999	−1.002 (0.316)	0.997	−1.614 (0.107)	0.991
	2001–2008	1.204 (0.229)	1.003	0.959 (0.338)	1.003	0.752 (0.452)	1.004	0.185 (0.853)	1.002
EP	1965–2008	0.312 (0.755)	1.001	0.138 (0.890)	1.001	0.589 (0.556)	1.004	−0.466 (0.641)	0.996
	1976–2008	0.295 (0.768)	1.002	−0.082 (0.935)	0.999	0.118 (0.906)	1.001	−0.889 (0.374)	0.991
	2001–2008	1.922 (0.055)	1.018	1.092 (0.275)	1.016	1.027 (0.304)	1.017	0.184 (0.854)	1.004
DE	1965–2008	2.117 (0.034)	1	1.935 (0.053)	1.001	1.523 (0.128)	1.003	1.650 (0.099)	1.005
	1976–2008	0.534 (0.593)	1	0.957 (0.339)	1.001	0.773 (0.439)	1.001	1.106 (0.269)	1.004
	2001–2008	0.352 (0.725)	1	0.815 (0.415)	1.001	1.377 (0.168)	1.003	1.457 (0.145)	1.009

Notes: h refers to the forecasting horizon. DM refers to the Diebold–Mariano statistic. The values within parantheses are the p-values.

Table 5: Out-of-sample results for FGLS versus OLS.

		$h = 1$		$h = 3$		$h = 6$		$h = 12$	
Predictor	Sample	DM	Theil U	DM	Theil U	DM	Theil U	DM	Theil U
DP	1965–2008	0.640 (0.522)	1	−0.180 (0.858)	1	−0.094 (0.925)	1	−1.353 (0.997)	0.997
	1976–2008	0.449 (0.654)	1	−0.813 (0.416)	0.998	−0.371 (0.710)	1	−1.661 (0.097)	0.996
	2001–2008	1.843 (0.065)	1.001	0.792 (0.429)	1.002	1.232 (0.218)	1	0.326 (0.744)	1.001
DY	1965–2008	−0.347 (0.729)	1	0.287 (0.774)	1	−0.565 (0.572)	0.999	−1.373 (0.170)	0.997
	1976–2008	−0.994 (0.320)	0.999	−0.455 (0.649)	0.999	−1.164 (0.245)	0.998	−1.665 (0.096)	0.995
	2001–2008	1.415 (0.157)	1.002	0.968 (0.333)	1.002	0.796 (0.426)	1.002	0.216 (0.829)	1.001
EP	1965–2008	1.105 (0.269)	1	−0.420 (0.675)	1	−0.229 (0.819)	1	−1.676 (0.094)	0.999
	1976–2008	0.959 (0.337)	1	−0.463 (0.643)	0.999	−0.197 (0.844)	1	−1.867 (0.062)	0.998
	2001–2008	1.585 (0.113)	1.002	0.840 (0.401)	1.003	0.773 (0.439)	1.001	−0.316 (0.752)	0.999
DE	1965–2008	2.643 (0.008)	1.001	2.001 (0.045)	1.001	1.414 (0.157)	1.001	1.393 (0.164)	1.002
	1976–2008	0.925 (0.355)	1	1.004 (0.316)	1	0.660 (0.510)	1	0.917 (0.359)	1.001
	2001–2008	0.808 (0.419)	1	0.712 (0.476)	1	1.202 (0.229)	1.001	1.461 (0.144)	1.003

Notes: See Table 4 for an explanation.

Table 6: Out-of-sample results for FGLS versus AOLS.

		$h = 1$		$h = 3$		$h = 6$		$h = 12$	
Predictor	Sample	DM	Theil U	DM	Theil U	DM	Theil U	DM	Theil U
DP	1965–2008	−2.486 (0.013)	0.989	−1.802 (0.072)	0.989	−1.530 (0.126)	0.988	−0.248 (0.804)	0.997
	1976–2008	−0.881 (0.378)	0.996	−0587 (0.557)	0.996	−0.297 (0.767)	0.997	0.581 (0.561)	1.007
	2001–2008	−2.119 (0.034)	0.985	−1.145 (0.157)	0.985	−1.247 (0.212)	0.98	−0.577 (0.564)	0.986
DY	1965–2008	−0.273 (0.785)	1	−0.576 (0.565)	1	−0.103 (0.918)	1	0.840 (0.401)	1.002
	1976–2008	1.069 (0.285)	1.001	0.515 (0.606)	1	0.825 (0.410)	1.002	1.533 (0.125)	1.005
	2001–2008	−2.119 (0.034)	0.985	−1.415 (0.157)	0.985	−1.247 (0.212)	0.98	−0.577 (0.564)	0.986
EP	1965–2008	−0.222 (0.824)	0.999	−0.249 (0.803)	0.999	−0.620 (0.535)	0.996	0.350 (0.726)	1.003
	1976–2008	−0.218 (0.827)	0.999	−0.002 (0.999)	1	−0.133 (0.894)	0.999	0.791 (0.429)	1.008
	2001–2008	−0.666 (0.506)	0.999	−0.309 (0.757)	1	−0.709 (0.478)	0.998	−0.157 (0.875)	0.999
DE	1965–2008	2.894 (0.004)	1.001	−1.580 (0.114)	1	−1.571 (0.116)	0.998	−1.802 (0.072)	0.997
	1976–2008	1.344 (0.179)	1	−0.841 (0.401)	1	−0.819 (0.413)	0.999	−1.212 (0.226)	0.998
	2001–2008	−1.966 (0.049)	0.985	−1.161 (0.246)	0.987	−1.042 (0.297)	0.984	−0.238 (0.812)	0.995

Notes: See Table 4 for an explanation.

Table 7: Utility gain from using FGLS.

	FGLS instead of AOLS				FGLS instead of OLS			
	$\phi = 3$		$\phi = 6$		$\phi = 3$		$\phi = 6$	
Predictor	$h = 1$	$h = 12$	$h = 1$	$h = 12$	$h = 1$	$h = 12$	$h = 1$	$h = 12$
1965–2008								
ED	0.580	1.123	0.580	1.123	0.707	0.377	0.657	1.401
DY	0.190	0.228	0.190	0.228	−0.499	−0.038	1.687	0.609
EP	0.445	0.701	0.445	0.701	1.922	0.178	0.381	0.702
DE	−0.007	0.014	−0.007	0.014	1.261	1.050	0.421	0.568
1976–2008								
ED	0.548	1.063	0.548	1.063	1.877	1.899	2.649	1.327
DY	0.166	0.235	0.166	0.235	−1.469	−5.780	1.176	5.092
EP	0.409	0.211	0.409	0.211	1.568	−0.051	3.008	6.266
DE	0.678	0.016	0.678	0.016	1.055	0.876	0.337	0.490
2001–2008								
ED	0.252	0.600	0.252	0.600	0.342	1.417	0.215	0.594
DY	0.079	0.181	0.079	0.181	−0.465	−2.276	0.248	2.046
EP	0.299	0.578	0.299	0.578	1.025	−0.198	1.899	0.501
DE	0.028	0.107	0.028	0.107	1.107	0.949	0.311	0.554

Notes: ϕ refers to the coefficient of relative risk aversion. See Table 4 for an explanation of the rest.

Table 8: Out-of-sample results for FGLS versus the historical average.

		$h = 1$		$h = 3$		$h = 6$	
Predictor	Sample	MSE–F	ENC–NEW	MSE–F	ENC–NEW	MSE–F	ENC–NEW
DP	1965–2008	17.193 (0.00)	62.400 (0.00)	37.570 (0.50)	2.078 (0.00)	1.664 (0.20)	6.670 (0.60)
	1976–2008	16.837 (0.00)	50.626 (0.00)	−32.239 (1.00)	0.216 (0.00)	5.104 (0.40)	6.840 (0.80)
	2001–2008	2.798 (0.00)	11.318 (0.00)	−6.924(0.70)	0.132 (0.00)	2.081 (0.56)	1.742 (0.56)
DY	1965–2008	34.144 (0.00)	38.114 (0.00)	1.728 (0.00)	0.754 (0.00)	−3.530 (0.10)	−0.643 (0.00)
	1976–2008	27.282 (0.00)	29.350 (0.00)	4.136 (0.00)	−0.868 (0.10)	−1.450 (0.40)	0.153 (0.00)
	2001–2008	5.639 (0.00)	6.318 (0.00)	0.754 (0.02)	0.575 (0.28)	0.095 (0.06)	0.280 (0.08)
EP	1965–2008	157.898 (1.00)	506.74 (1.00)	31.523 (0.80)	2.583 (0.00)	−3.389 (0.20)	1.117 (0.40)
	1976–2008	−11.898 (0.20)	260.01 (1.00)	−30.970 (0.90)	−5.275 (0.00)	−3.879 (0.50)	0.319 (0.00)
	2001–2008	−57.894 (1.00)	10.641 (0.98)	−13.506 (0.98)	−4.727 (0.48)	−2.785 (0.72)	−0.962 (0.40)
DE	1965–2008	2.728 (0.20)	12.080 (0.80)	2.801 (0.70)	−0.133 (0.10)	3.420 (0.60)	−0.795 (0.40)
	1976–2008	10.919 (0.90)	16.047 (1.00)	1.665 (0.40)	1.758 (0.50)	0.458 (0.00)	0.848 (0.80)
	2001–2008	9.557 (0.94)	11.474 (1.00)	2.312 (0.82)	1.404 (0.86)	0.274 (0.24)	0.269 (0.44)

Notes: The values within parantheses are the bootstrapped p-values. See Table 4 for an explanation of the rest.

Table 9: Out-of-sample results for $h = 1$ when FGLS is implemented using $\rho_0 = 0.9999$.

		Historical average		OLS		AOLS	
Predictor	Sample	DM	Theil U	DM	Theil U	DM	Theil U
DP	1965–2008	−2.222 (0.00)	0.938 (0.00)	−2.261 (0.00)	0.939 (0.00)	−2.062 (0.00)	1.000 (1.00)
	1976–2008	−2.151 (0.00)	0.925 (0.00)	−2.140 (0.00)	0.928 (1.00)	−2.143 (0.00)	1.000 (1.00)
	2001–2008	−0.770 (0.00)	0.942 (0.00)	−0.801 (0.00)	0.941 (1.00)	−0.763 (0.00)	1.000 (1.00)
DY	1965–2008	−1.533 (0.00)	0.986 (1.00)	−1/578 (0.00)	1.000 (1.00)	−1.557 (0.00)	1.000 (1.00)
	1976–2008	−2.070 (0.50)	0.979 (1.00)	−1.966 (1.00)	0.983 (1.00)	−2.110 (0.50)	1.000 (1.00)
	2001–2008	−0.146 (0.00)	0.997 (1.00)	0.243 (0.00)	0.996 (1.00)	−0.213 (0.00)	1.000 (1.00)
EP	1965–2008	−1.672 (0.00)	0.962 (0.00)	−1.718 (0.00)	0.963 (0.00)	−1.760 (0.00)	1.000 (1.00)
	1976–2008	−1.935 (0.00)	0.947 (0.00)	−1.983 (0.00)	0.950 (0.00)	−2.033 (0.00)	1.000 (1.00)
	2001–2008	−2.434 (0.00)	0.902 (0.00)	−3.073 (0.00)	0.888 (0.00)	−2.748 (0.00)	1.000 (1.00)
DE	1965–2008	−0.196 (0.00)	0.999 (0.00)	0.994 (0.50)	1.004 (1.00)	1.042 (0.50)	1.000 (1.00)
	1976–2008	0.927 (1.00)	1.005 (1.00)	0.956 (0.50)	1.005 (1.00)	0.951 (0.50)	1.000 (1.00)
	2001–2008	1.666 (1.00)	1.021 (1.00)	1.757 (1.00)	1.024 (1.00)	1.752 (1.00)	1.000 (1.00)

Notes: See Tables 4 and 8 for an explanation.