Predictive Regressions in Predictive Systems*

Tamás Kiss†

October 25, 2018

Abstract

This paper analyses predictive regressions in a predictive system framework, where the predictor is an imperfect proxy for the expected returns. I show that when there are differences between the dynamic structure of the expected return and the predictor, the predictive regression uses predictive information inefficiently, resulting in inferior model fit. The effect is especially strong if the predictors and the expected return are highly, but not equally, persistent. As a solution, I propose a persistence adjustment for the predictive regression. The resulting estimator is a two-stage method, where the expected return and predictor processes are modelled separately, allowing for each to have distinct dynamic properties. Simulations, as well as empirical results, show that the method leads to both better in-sample fit and real-time forecasting performance. The empirical results highlight that the proposed method is especially useful in the case of multiple predictors.

JEL classification: C22, G1.

Keywords: Persistence adjustment; Predictive system; Return predictability;

*I have benefited from comments by Erik Hjalmarsson, Adam Farago, Hossein Ashgarian, Joakim Westerlund, Emre Aylar and the participants at the 5th Annual PhD Workshop (2018) at the University of Gothenburg, and the joint KWC-CFF workshop (2018) in Varberg.

†Department of Economics, Centre for Finance, University of Gothenburg. Contact information: P.O. Box 640, SE 405 30 Gothenburg, Sweden. Email: tamas.kiss@economics.gu.se.

1 Introduction

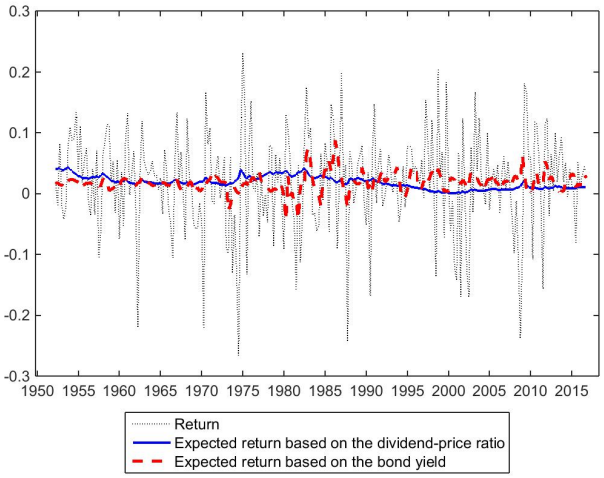
Since the seminal contribution of Campbell and Shiller (1988), several papers have argued for the existence of time-varying expected returns (Lettau and Ludvigson, 2001; Ang and Bekaert, 2007; Cochrane, 2008, 2011). The consensus in the financial literature has subsequently converged towards accepting that return predictability probably exists, and the focus has shifted towards understanding how potential predictors contribute to predictability. In parallel, the large empirical literature has found that the evidence on predictability using predictive regressions can be subject to statistical problems (Goyal and Welch, 2008; Koijen and Van Nieuwerburgh, 2011), which has spurred the development of sophisticated inference techniques for testing the null of no return predictability. The proposed tests primarily deal with correcting for the persistent regressor bias in order to conduct valid tests of whether returns are predictable (Cavanagh et al., 1995; Stambaugh, 1999; Lewellen, 2004; Torous et al., 2004; Campbell and Yogo, 2006; Jansson and Moreira, 2006; Kostakis et al., 2015).

The inferential problem changes, however, when the aim is to assess which variables are useful predictors, rather than explicitly test a null of no predictability. In this case, the key is to understand how a certain predictor (or predictors) is related to future expected returns, and how this relationship can best be estimated. Clearly, all predictive regressions cannot *simultaneously* be the true data-generating process for the expected return. For instance, univariate regressions with valuation ratios and term structure variables imply expected return processes with fairly different properties. Both types of regressions can still be useful for understanding predictability, as these variables most likely carry information about future expected returns. However, they most probably do so imperfectly in the sense that predictors only proxy for the expected return series, as described by the predictive system in Pástor and Stambaugh (2009). That is, expected return variation is only partially recovered in any given specification.

The current paper studies predictive regressions in the presence of predictor imperfection. I look at two, non-mutually exclusive, forms of imperfection. First, predictors might not explain the full variation in expected returns in the sense that the latent expected

return process is not a linear combination of the predictor variables. This form of imperfection essentially reflects a fundamental lack of information in the predictive regression formulation and cannot be fully controlled for within the model. Second, predictors and the expected returns might have different dynamic properties. I focus on this latter form of imperfection, which can be controlled for within the predictive system. I show that based on the standard predictive regression the explanatory power of the predictor decreases as the difference between the persistence of the expected return and the predictor grows. This effect is particularly strong if the variables are highly persistent. In the limit, where both the predictor and the expected return are (nearly) non-stationary, the predictive regression becomes spurious (similar to the problems described in Ferson et al., 2003; Deng, 2013).

Figure 1: **Implied expected return processes from predictive regressions**



Notes: The figure shows realized excess returns of the CRSP value weighted index (dotted line) and expected returns implied by running univariate predictive regressions $r_{t+1} = \alpha + \beta x_t + e_{t+1}$, where x_t is either the dividend-price ratio (solid line) or the (detrended) yield on the long term government bond (dashed line). The sample runs between 1952 and 2016. Further details on the variables are provided in Section 5.

To intuitively understand why differences in the time-series structure are important, consider the simple example in Figure 1, where expected returns are calculated using univariate predictive regressions based on two different predictors: the dividend-price ratio and the (detrended) long-term bond yield.[1] Unsurprisingly, the two expected return

[1]Detailed description of the specifications can be found in the description of Figure 1.

series are markedly different from each other, in particular in terms of their dynamic properties. Figure 1 thus indicates that information on the persistence of the expected return can be useful when estimating the effect of the predictors.

To pursue this idea, I propose a persistence adjustment to the predictive regression. Incorporating the assumption that expected returns follow a first order autoregressive process, the Persistence Adjusted Predictive Regression (PAPR) improves upon the standard ordinary least squares (OLS) estimation in terms of model fit and real-time forecasting performance. The gain in explanatory power comes from the fact that the persistence adjustment disconnects the time-series dynamics of the predictor(s) from the persistence of expected returns. The persistence adjustment is operationalized by a two-step estimation framework. In the first step, the parameters governing the dynamics of the predictors are calculated using standard least squares technique. In the second step, the latent expected return process is obtained by minimizing the variance of the unexplained returns. The method belongs to the class of extremum estimators described by, for example, Newey and McFadden (1994), and hence its properties are well-known. In particular, the standard errors can be calculated in a straightforward way, accounting for the two-step nature of the estimation procedure.

The predictive system is formally represented as a state-space model. In the general case, estimation of the expected return process can be performed by the Kalman Filter. Asymptotically this yields optimal expected return estimates, connecting the variation in expected returns to the predictor and/or to past realized returns. I show that the PAPR is a restricted version of the Kalman Filter. The PAPR makes use of information in the predictive variables, but does not connect expected return variation to realized returns. It thus provides the optimal expected return series given the information in the predictor, but ignores information in past returns. The upside is that it requires less parameters to be estimated than the Kalman Filter, which translates into less parameter uncertainty, and better out-of-sample forecasts.[2] In the special case where the predictor

[2]The information loss in the PAPR, relative to the Kalman Filter actually appears quite small. When the true parameters of the model are assumed known (i.e., no parameter uncertainty), the advantage of the Kalman-filtered expected returns is not particularly large.

and the expected return have the same time-series dynamics, the PAPR collapses to a standard predictive regression estimated by OLS. Thus, from a practical perspective, the proposed persistence adjustment connects the structural assumptions of the state space model with the estimation framework of the predictive regression.

The performance of the different specifications of expected returns are compared through simulations and an empirical application. A Monte Carlo experiment shows that the PAPR outperforms both OLS and the Kalman Filter in terms of real-time forecasting performance, which suggests that the effect of ignoring past return information is dominated by the reduced parameter uncertainty. The results also indicate that the advantage of the PAPR over OLS increases as the differences in the dynamics grow, which is in line with the theoretical discussion.

The empirical analysis is based on quarterly excess stock market returns and the three predictors used in Pástor and Stambaugh (2009): the dividend-price ratio, the consumption-to-wealth ratio (cay) by Lettau and Ludvigson (2001) and the detrended yield on the 30-year US government bond. Results confirm that the persistence adjustment involves a bias-variance trade-off compared to least squares estimation of the predictive regression. Since more parameters are estimated using the same amount of information, the parameter estimates of the PAPR tend to have larger standard errors. The advantage of the persistence adjustment comes from the fact that the time-series dynamics of the expected return are estimated separately from the predictors. In particular, when the predictor has a relatively low persistence (as in the case of the univariate regression using the bond yield as a predictor), using the PAPR is useful because it can capture the potentially higher persistence of the expected return. This becomes even clearer in the case of several predictors, where the PAPR outperforms least squares both in-sample and out-of-sample.

The remainder of the paper is organised as follows. Section two discusses the model and the properties of least squares estimation in the predictive system. Section three describes the PAPR and its relationship to other estimation methods. Section four presents Monte Carlo simulations analysing the properties of the PAPR. Section five presents an empirical

application of the method and Section six concludes. The appendix contains technical derivations and supplementary results.

2 Predictive regressions and predictive systems

The workhorse model of empirical research on return predictability is the predictive regression,

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

where r_{t+1} is an observed excess return series (usually stock market index returns in excess of a risk free rate) and x_t is a predictive variable.[3] This specification implies $E_t(r_{t+1}) = \alpha + \beta x_t$, that is, the conditional expected return is a linear function of the predictive variable. In particular, the predictive regression implies that the dynamics of the expected return is identical to the dynamics of the predictor, otherwise the regression is misspecified. The key advantage of this model is that it is straightforward to estimate using least squares, and standard testing procedures (potentially corrected for the persistent regressor bias described in Stambaugh, 1999) are readily available. Therefore, it is a simple and well-understood tool to decide whether certain variables predict excess returns. Many predictors have been proposed and tested in the literature, both in univariate settings and in combinations (see, for example, Goyal and Welch, 2008 and the references therein).

Pástor and Stambaugh (2009) introduce the predictive system, where the predictors are *not* perfect proxies of the expected returns. It is a convenient framework to analyse the case when the time-series dynamics of expected returns and the predictor differ, since it allows the dynamics of the expected return series to be defined separately. Formally,

[3] In the theoretical discussion, I consider the univariate case only. Results straightforwardly extend to the multiple predictor case, unless it is discussed separately (as in Appendix C).

the predictive system is given by the following state space model,

$$r_{t+1} = \mu_t + u_{t+1}, \tag{2}$$

$$\mu_{t+1} = (1-\gamma_\mu)\bar{\mu} + \gamma_\mu \mu_t + w_{t+1}, \tag{3}$$

$$x_{t+1} = (1-\gamma_x)\bar{x} + \gamma_x x_t + \epsilon_{t+1}. \tag{4}$$

r_{t+1} and x_t are the same as in case of the predictive regression specification, and $\mu_t = E_t(r_{t+1})$ is the conditional expected return process, modelled separately. The innovation processes $\{u_t, w_t, \epsilon_t\}_{t=0}^{\infty}$ are assumed to be zero mean, serially independent martingale difference sequences with a finite covariance matrix. In this specification, the autoregressive parameters of the expected return (γ_μ) and the predictor process (γ_x) need not coincide. The correlation between the innovations of the expected return and the predictor, $\rho_{w\epsilon} = Corr(w_t, \epsilon_t)$, determines the informativeness of the predictive variable. $\rho_{w\epsilon} = 0$ implies that the predictor variable is completely uninformative. In the other extreme, $\rho_{w\epsilon} = 1$, together with $\gamma_x = \gamma_\mu$, implies that expected returns are completely pinned down by the predictor. In this case, the system reduces to the predictive regression in equation (1), with equation (4) describing the evolution of the predictor.

Under the assumption that returns and the predictor are generated by equations (2), (3) and (4), the properties of the predictive regression in equation (1) can be derived. In particular, if we assume stationarity in the system ($\gamma_x < 1,\ \gamma_\mu < 1$), the OLS estimator of the slope coefficient in the predictive regression satisfies the standard result,

$$\hat{\beta}_{OLS} \xrightarrow{p} \frac{E\left((r_{t+1}-\bar{r})(x_t-\bar{x})\right)}{E\left((x_t-\bar{x})^2\right)} = b\frac{1-\gamma_x^2}{1-\gamma_\mu\gamma_x}, \tag{5}$$

where $b = \rho_{w\epsilon}\frac{\sigma_w}{\sigma_\epsilon}$ is the coefficient determining the relationship between the expected return and predictor innovations (hereafter referred to as the innovation slope coefficient). The formula shows that the slope coefficient of the OLS estimator depends on the relationship between the innovations and the differences in persistence. To analyse this expression further, I fix the amount of predictability, or more specifically the ratio of expected to unexpected return variation. Define the quantity $\eta = \sigma_w \big/ (\sigma_u\sqrt{1-\gamma_\mu^2})$, gov-

erning the amount of predictability present in returns (also referred to as normalized beta, for example, in Wachter and Warusawitharana, 2009, 2015; Lucivjanska, 2018).[4] Using this notation, the slope coefficient of the predictive regression can be decomposed into three parts and a scale factor,

$$\hat{\beta}_{OLS} \xrightarrow{p} \eta\rho_{w\epsilon}\frac{\sqrt{1-\gamma_\mu^2}\sqrt{1-\gamma_x^2}}{1-\gamma_x\gamma_\mu}\frac{\sigma_u\sqrt{1-\gamma_x^2}}{\sigma_\epsilon} \equiv \beta_{OLS}^{plim}. \tag{7}$$

First, the asymptotic limit of the OLS estimator depends positively on the relative variation of the expected return, η. The intuition is straightforward: the larger the amount of predictability, the stronger the regression evidence becomes. Second, β_{OLS}^{plim} depends on the correlation between the predictor and the expected return ($\rho_{w\epsilon}$), as a better proxy for the predictor implies a larger slope coefficient in the predictive regression. The third component highlights the importance of distinguishing between the time-series properties of the expected return and the predictor series. The value of this term, which depends only on the persistence parameters, is between zero and one, and it is equal to one only if $\gamma_\mu = \gamma_x$. The strongest predictive relationship can thus be detected if the persistence of the predictor and the expected return are aligned.[5] The first two components (the amount of expected return variation η and the correlation $\rho_{w\epsilon}$) are "fundamental"quantities of the model, which directly determine the amount of variation a predictor can explain. Without any further information, these quantities must be viewed as given and fixed. In contrast, the difference in persistence is a feature that can be corrected for by using the structural assumptions of the model, as discussed further in Section 3.

Note that if $\gamma_x \to 1$ while $\gamma_\mu < 1$, the OLS estimator converges to zero keeping other parameters, in particular the scaling and the degree of predictability, constant. It is in line with the fact that a non-stationary variable cannot be used to capture stationary variation. The same result holds if $\gamma_\mu \to 1$ and $\gamma_x < 1$, since, analogously, a stationary

[4]Using the quantity η the amount of explained return variance can be rewritten as

$$R_{true}^2 = \frac{\eta^2}{1+\eta^2} \tag{6}$$

Note that η is unobservable, since it depends on the parameters of the latent expected return process.

[5]Similar results for the regression t-statistics and the R^2 are derived in Appendix A.

variable cannot capture the variation in a non-stationary variable. Furthermore, if both of the persistence parameters approach one, the limit is not well defined. In particular, the limit

$$\lim_{(\gamma_x,\gamma_\mu)\to(1,1)} \frac{\sqrt{1-\gamma_\mu^2}\sqrt{1-\gamma_x^2}}{1-\gamma_x\gamma_\mu}$$

depends on the relative rates of convergence for γ_x and γ_μ. The special case, when both the expected return and the predictor approach the non-stationary region therefore needs to be analyzed separately. This case is of interest both because of the large literature on the effect of persistent regressor bias in predictive regressions (Stambaugh, 1999; Lewellen, 2004; Campbell and Yogo, 2006; Phillips, 2014, among others), and the empirical fact that many of the important predictors (in particular, valuation ratios) exhibit high persistence. The full formal analysis is relegated to Appendix B, but the main finding is that the correlation between expected return and predictor innovations plays a crucial role when predictors are nearly non-stationary. If the correlation is not strong, the regression t-statistic is dominated by the spurious regression effect, making inference invalid. In fact, the spurious predictive regression literature (Ferson et al., 2003; Deng, 2013), where the predictor is completely uninformative about expected returns, is a special case of the results derived in Appendix B. On the other hand, when the correlation between the innovations is high, the difference in persistence does not enter the asymptotic distribution of the test statistics. In this case, the predictor and the expected return become asymptotically equivalent. In a knife-edge case however, the difference in persistence does play a role, affecting the distribution of the t-statistic through an extra term that enters due to imperfection.

Overall, the predictive system in the (near) non-stationary case becomes tenuous, where meaningful inference is only possible in very specific cases. That is, unless the data-generating process is in the knife-edge case described in Proposition 1 in Appendix B, the predictive system either results in spurious predictability or asymptotically reduces to the predictive regression. In the remainder of the analysis, I therefore focus on the stationary case, where the predictive system does not collapse to either of these special cases.

3 Inference under imperfect predictors

3.1 Persistence Adjusted Predictive Regression

As discussed in the previous section, the predictive regression is misspecified if the predictor and the expected returns have different persistence. In this section, I propose a Persistence Adjusted Predictive Regression (PAPR), which is a method that explicitly corrects the predictive regression to account for the difference in the persistence of the predictor and the expected returns.

Assume that the data-generating process is described by the predictive system in equations (2), (3) and (4). Given the parameters of the model, the innovations of the predictor, $\{\epsilon_t\}_{t=1}^{T}$, can be calculated by applying the dynamics in equation (4). The expected return process is formed from the innovations w_t, and although these are unobservable, a projected expected return series can be calculated using the predictor innovations and the parameters of the model. In particular, the least squares projection of the expected return innovation is given by $w_t|\epsilon_t = b\epsilon_t$, where b is the innovation slope coefficient introduced in equation (5). The projected expected return series can then be calculated as

$$\mu_t = \bar{\mu} + \sum_{s=1}^{t} \gamma_{\mu}^{t-s} b\epsilon_s. \tag{8}$$

That is, the projected innovations $b\epsilon_s$ are used in the autoregressive filter governing the dynamics of the expected return process. Note that if $\gamma_\mu = \gamma_x$, the expected return series implied by equation (8) reduce to

$$\mu_t = \bar{\mu} + \sum_{s=1}^{t} \gamma_{x}^{t-s} b\epsilon_s = \bar{\mu} + b\sum_{s=1}^{t} \gamma_{x}^{t-s} \epsilon_s = \alpha + \beta x_t,$$

in which case the expected return process implied by the projection is identical to that of the predictive regression. Estimating (8) with γ_μ as a free parameter is essentially an augmented version of OLS estimation, where the potentially different persistence of the predictor and the expected return is taken into account.

Parameter estimation of the PAPR can be performed by minimizing the forecast error.

The objective function can be written as

$$Q(\theta) = \frac{1}{T-1} \sum_{t=1}^{T-1} \left(r_{t+1} - \tilde{\mu}_t(\theta) \right)^2, \tag{9}$$

where θ denotes all parameters of the model. The structure of the problem suggests that the estimation can be done in two steps. In particular, let $\theta = (\theta_1, \theta_2)$, where $\theta_1 = \{\bar{x}, \gamma_x\}$ includes the parameters of the predictor process and $\theta_2 = \{\bar{\mu}, \gamma_\mu, b\}$ contains the parameters of the expected return process and the innovation slope coefficient. Since the predictor follows a simple autoregression, its parameters can be efficiently estimated by OLS, and its innovations can be calculated in the first step. In the second step, the objective function in (9) can be minimized with respect to θ_2 to get the estimates of the parameters of the expected return process and the innovation slope coefficient,

$$\begin{aligned}
\hat{\theta}_2 &= \arg\min_{\theta_2} Q(\hat{\theta}_1, \theta_2) \\
&= \arg\min_{\theta_2} \frac{1}{T-1} \sum_{t=1}^{T-1} \left(r_{t+1} - \tilde{\mu}_t(\hat{\theta}_1, \theta_2) \right)^2 \\
&= \arg\min_{\theta_2} \frac{1}{T-1} \sum_{t=1}^{T-1} \left(r_{t+1} - \bar{\mu} - \sum_{s=1}^{t} \gamma_\mu^{t-s} b \hat{\epsilon}_s \right)^2 \\
&= \arg\min_{\theta_2} \frac{1}{T-1} \sum_{t=1}^{T-1} \left(r_{t+1} - \bar{\mu} - \sum_{s=1}^{t} \gamma_\mu^{t-s} b \left[x_s - (1-\hat{\gamma}_x)\hat{\bar{x}} - \hat{\gamma}_x x_{s-1} \right] \right)^2,
\end{aligned}$$

where $\hat{\epsilon}_t$ is the fitted residual of the predictor.

Two-step estimators constitute a special case of extremum estimators, hence their asymptotic properties are well known (Newey and McFadden, 1994). In particular, since the parameters of the predictor process can be consistently estimated by OLS, the estimates in the second step can also be consistently estimated. The asymptotic normality of the estimator has the usual form, except that standard errors of the estimates in the second step need to take into account the estimation error of the first step (see Appendix C).

Analogous to classical regressions, the PAPR can also be easily extended to the multivariate regression case. If the variables $x_t^1, x_t^2, \ldots x_t^J$ are all potential predictors of the

expected return, the first-step innovation series $\hat{\epsilon}_t^1, \hat{\epsilon}_t^2, \ldots \hat{\epsilon}_t^J$ are obtained using a multivariate time-series model for the predictors. The expected return projection is then formulated as $w_t|\epsilon_t^1, \epsilon_t^2, \ldots, \epsilon_t^J = \sum_{j=1}^J b_j \epsilon_t^j$. The second step is modified such that all the innovation slope coefficients $\{b_j\}_{j=1}^J$ are jointly estimated with the parameters of the expected return process.

Another assumption of the PAPR that can be easily relaxed is the time-series dynamics of the predictor. Equation (4) can be redefined using a more general time series model, and the predictor innovations are obtained by estimation of the defined model. Asymptotic results and further discussion on the implementation of the two-step procedure are found in Appendix C.

3.2 Comparison to Kalman Filter

Pástor and Stambaugh (2009) use a Kalman Filter to estimate the predictive system. Since their data-generating process is identical to the one proposed in this paper, I compare the persistence adjustment to the Kalman Filter estimation of the system. In particular, the key difference between the methods is that the PAPR contains information only from the predictor (its covariance structure and cross-correlation with the returns), while the Kalman Filter connects the expected return variation not only to past predictor innovations but also to past returns. To see this, consider the regression formulation of the conditional expected return in the state space model (the derivation can be found in Pástor and Stambaugh, 2009),

$$\mu_t = \bar{\mu} + \sum_{s=1}^{t} \omega_s (r_s - \bar{\mu}) + \sum_{s=1}^{t} \delta_s \epsilon_s. \tag{10}$$

The parameters of the linear model, $\omega_s = m(\gamma_\mu - m)^{t-s}$ and $\delta_s = n(\gamma_\mu - m)^{t-s}$ depend on the persistence of the expected returns and the parameters m and n, which, in turn, are functions of the parameters in equation (2)-(4) and the covariance matrix of the error terms. m and n are parameters measuring how much (past) returns and predictors contribute to the expected return variation, respectively.[6] In general, these parameters

[6]Their exact dependence on the parameters of the underlying data-generating process is given by

need to be estimated.

Kalman Filter estimation can be viewed as estimating the parameters of equation (10), without imposing any further assumption on the parameters. In contrast, using the PAPR is equivalent to imposing $m = 0$. In this case $\omega_s = 0$, $\delta_s = \gamma_\mu^{t-s}$ and $b = n$.[7] Thus, equation (10) collapses to the specification of the PAPR in equation (8). Setting $m = 0$ is an assumption, forcing past returns to have no effect on the expected return prediction. So, while the Kalman Filter attributes time variation in expected returns to both past returns and predictor innovations, the proposed two-step method shuts down the channel through which past returns directly operate.

If the model is correctly specified, the Kalman Filter estimated by maximum likelihood results in an asymptotically optimal estimate of the expected return. However, the typical sample size in the current return predictability setting is relatively small compared to the number of parameters that need to be estimated in a full state-space model. Therefore, the parameter uncertainty is potentially large in the Kalman Filter estimation, and the asymptotic optimality results might not be relevant in empirically occurring sample sizes. The advantage of the PAPR is that it reduces the parameter uncertainty compared to the Kalman Filter. That is, the (asymptotic) bias caused by imposing the restriction $m = 0$ is traded-off against the reduced number of parameters. The PAPR can thus be viewed as a more robust way of estimating expected returns, while still taking into account the fact that the persistence of the predictor and the expected return are potentially different.

In the following two sections, I analyse the PAPR and further compare it with the Kalman Filter and OLS both in Monte Carlo simulations and in an empirical application. Essentially, I compare three different specifications of the expected return: $\alpha + \beta x_t$ for the standard predictive regression, equation (8) for the PAPR and equation (10) for the Kalman Filter. I focus on their performance both in-sample (how well they describe expected returns) and out-of-sample (how they perform in terms of real-time forecasting).

equations (A36) and (A37) in Pástor and Stambaugh (2009).

[7] $n = (\sigma_{w\epsilon} - m\sigma_{u\epsilon})\sigma_\epsilon^{-2}$ in the general formulation in Pástor and Stambaugh (2009). $m = 0$ corresponds to $n = \sigma_{w\epsilon}\sigma_\epsilon^{-2} = b$.

4 Simulations

In this section I perform a Monte Carlo simulation to assess the properties of the Persistence Adjusted Predictive Regression (PAPR) in a predictive system. I present two sets of simulation results, which closely relate to the theoretical discussion in Sections 2 and 3. First, I show that the predictive regression cannot capture the persistence of the expected return, when it is separate from the predictor. Therefore, the PAPR is able to produce a better estimate of the expected return (a higher in-sample fit), since it estimates the persistence parameter of the expected return separately. Second, I carry out an analysis of real-time forecasting performance comparing the predictive regression, the PAPR and the Kalman Filter estimation of the full system.

4.1 Simulation setup

All simulations are based on the assumption that the data-generating process is given by the predictive system described by equations (2)-(4), where the innovations follow a jointly normal process. The baseline parametrization of the system is as follows. $\bar{\mu} = 0.018$ and $\bar{x} = 0.03$ are the unconditional means of the return and the predictor, respectively. These values correspond to the quarterly unconditional mean of the excess return and the dividend-price ratio. Expected returns are assumed to explain five percent of total return variation ($\eta^2 = 0.05$). The default value for the persistence of the expected return is $\gamma_\mu = 0.9$ and the autoregressive parameter of the predictor $\gamma_x \in [0.5, 0.99]$ is specified for each simulation.

The standard deviations of the unexpected and expected returns are set such that the quarterly unconditional volatility is eight percent. Given the parameters (in particular, η and γ_μ) above, this implies $\sigma_u = 0.081$, and $\sigma_w = 0.011$. Further, using the value of γ_x, the standard deviation of the predictor, σ_ϵ is calculated to ensure that $\beta_{OLS}^{plim} = \rho_{w\epsilon}$. This choice makes the comparison over specifications easier, since it imposes the same asymptotic limit for the OLS estimator in each specification.

The correlation structure of the innovations is chosen to reflect the presence of imperfection. The default value for predictor imperfection, i.e., the correlation between the

expected return and predictor innovation is set to $\rho_{w\epsilon} = 0.9$. The correlation between expected and unexpected returns is set to $\rho_{uw} = -0.7$ to capture the negative correlation for the dividend-price ratio that has been documented in the literature. It is also assumed that $\rho_{u\epsilon} = \rho_{uw}\rho_{w\epsilon}$, which implies that unexpected returns and the predictor are only correlated through their correlation with the expected return innovations. All simulations are performed with a sample size $T = 200$, and the results are based on 1000 repetitions in each case. These parameter choices are kept throughout the simulations, unless otherwise noted.

4.2 In-sample results

The first set of simulation results highlights the misspecification in the predictive regression. Results are obtained by fixing all parameters at default values (in particular, $\gamma_\mu = 0.9$) and varying the autoregressive coefficient of the predictor between $\gamma_x = 0.5$ and $\gamma_x = 0.99$. Table 1 shows summary statistics of estimation results for the predictive regression and for the PAPR. The first observation is that the slope coefficient estimates of the standard predictive regression (first column in Panel a) are relatively close to the true underlying value (parameters are set such that the slope coefficient of the predictive regression is $\beta = 0.9$).[8] However, the expected return implied by the standard predictive regression is $\hat{\mu}_{t,OLS} = \hat{\beta}x_t$, so its persistence is pinned down by x_t. Therefore the estimated persistence of the expected return is biased, unless $\gamma_\mu = \gamma_x$ (as seen in the third column of Table 1, Panel a). The advantage of the persistence adjustment comes from the fact that it can estimate γ_μ with less bias. In particular, the implied regression slope coefficient $\hat{\beta}_{PAPR,implied} = \hat{b}\frac{1-\hat{\gamma}_x^2}{1-\hat{\gamma}_\mu\hat{\gamma}_x}$ in the first column of Panel (b) is similar to those of the standard predictive regression. However, the persistence of the estimated expected returns no longer depends on γ_x (third column of Table 1, Panel b). [9]

[Table 1 about here.]

[8] The only exception is for the case where the persistence of the predictor is very high, suggesting that in this case the usual inference is not valid, as described in the theoretical section.

[9] All the estimates of the autoregressive parameters are downward biased due to the small sample bias present in OLS estimation. Nevertheless, this does not influence the comparison between the standard predictive regression estimated by OLS and the PAPR.

The ability of the PAPR to capture the difference in persistence translates into better model fit. To illustrate this, I calculate the in-sample R^2 of the models. It measures how much return variation a given model explains and is defined as

$$IS - R^2 = 1 - \frac{Var(r_{t+1} - \hat{\mu}_t)}{Var(r_{t+1})}, \tag{11}$$

where $\hat{\mu}_t$ is the expected return generated by the model. In case of OLS, the measure is identical to the usual R^2, while in a non-linear specification it is usually referred to as the pseudo-R^2. Figure 2 shows that the PAPR is better than the standard predictive regression in terms of in-sample R^2, and its advantage increases as the difference in persistence grows, confirming the theoretical results in Section 2.

[Figure 1 about here.]

4.3 Out-of-sample results

The second set of simulations analyses the PAPR in terms of its ability to predict expected returns in a real-time forecasting setup, and compares it to that of the predictive regression and the estimation of the full state-space system by the Kalman Filter. The real-time forecasting performance of the model is measured by its out-of-sample R^2 defined by Goyal and Welch (2008),

$$OOS - R^2 = \frac{MSFE_{benchmark} - MSFE_{model}}{MSFE_{benchmark}}, \tag{12}$$

where $MSFE_{model}$ ($MSFE_{benchmark}$) is the mean squared forecasting error of the model (benchmark). The historical mean forecast (that is, $\mu_t = \frac{1}{t}\sum_{s=1}^{t} r_s$) is used as the benchmark model. A positive $OOS - R^2$ therefore implies that the model outperforms the constant expected return model. The training sample is always set equal to 200 observations and the simulations are based on 1000 one period ahead forecasts of the expected return. Out-of-sample R^2 values are calculated for the default parametrization and γ_x is varied between 0.5 and 0.99.

[Figure 2 about here.]

Figure 3 shows the results of the simulations. They imply that the PAPR typically outperforms both the standard OLS estimation and the Kalman Filter. Since the PAPR involves more parameters, its advantage over the standard predictive regression is smaller if there is no large difference in persistence. However, provided even a relatively small difference in persistence, PAPR outperforms the standard predictive regression. The results in Figure 3 further show that the maximum likelihood estimation of the Kalman Filter is not well suited for out-of-sample forecasts due to parameter uncertainty. It always underperforms the other methods but also the historical mean specification. This confirms the results in Lucivjanska (2018), showing that the predictive regression is usually better in terms of out-of-sample performance.

[Figure 3 about here.]

To illustrate that the weak performance of the Kalman Filter is due to estimation uncertainty, Figure 4 presents results for the case where the parameters of the expected return process are known. That is, there is no estimation error, and the differences in the models are entirely due to how expected returns are calculated ($\alpha + \beta x_t$ for the predictive regression, equation (8) for the PAPR and the filtering equations described by Pástor and Stambaugh (2009) for the Kalman Filter). In this empirically infeasible case, the Kalman Filter provides optimal expected return series. It is reflected in Figure 4, with the Kalman Filter generating the highest out-of-sample R^2. However, the figure also highlights the importance of adjusting for the difference in persistence. The prediction made by the standard predictive regression is dominated by the PAPR, which, in turn, is remarkably close to the Kalman Filter. The results in Figure 4 thus suggest that the advantage of the full system estimation is limited, given the similarity between PAPR and the Kalman Filter.

5 Empirical analysis

I now turn to an empirical analysis using the PAPR method described above. I estimate various models to predict the quarterly returns on the Center for Research in Security Prices (CRSP) value-weighted stock market index between 1952 and 2016. Excess returns are calculated using the 30-day Treasury bill as the risk-free rate. The predictors are the same as those considered in Pástor and Stambaugh (2009). The dividend-price ratio (dp) is calculated using returns on the CRSP value-weighted index with and without dividends. The consumption-to-wealth ratio (cay) is obtained from Martin Lettau's website. The bond yield (by) variable is the difference between the 30-year government bond yield and its twelve month moving average in the CRSP Treasuries file.

Descriptive statistics for the variables are shown in Table 2. The first order autocorrelations in the third column show that the predictors are quite different in terms of their time-series properties. The autoregressive parameter for the dividend-price ratio is 0.97, which implies a fairly high persistence, close to non-stationarity. On the other hand, the first order autocorrelation of the bond yield is only 0.61, implying a relatively fast mean reversion. The cay variable is in the middle with an autoregressive parameter of 0.82. These numbers also suggest that the expected return processes implied by the univariate regressions are likely quite different from each other.

[Table 2 about here.]

Panel (a) in Table 3 presents results from univariate OLS regressions and a multivariate regression including all the variables. Results show that, in this dataset, the dividend-price ratio is the weakest predictor of the expected return, while the other two variables exhibit stronger relationships with one-quarter ahead returns. Including all of these variables in the regression leaves the coefficient on each variable largely unchanged, which suggests that the three predictor variables convey different information, and multicollinearity is not particularly large.

[Table 3 about here.]

The PAPR results show similar patterns. Innovations in the first step are obtained through a first order autoregressive filter (shown in Table 4).[10] The estimates of the second step, i.e., the estimated persistence of the expected return and the innovation slope coefficients, are shown in panel (b) of Table 3. Overall, the univariate results of the PAPR are similar to those of the standard regression estimated by OLS. The autocorrelation coefficient of the expected return (γ_μ) is estimated with a fairly high precision. The innovation slope coefficients are estimated with higher standard errors than the corresponding OLS slope coefficients, reflecting that in case of the PAPR there are more parameters to estimate compared to the standard predictive regression. When all three variables are included in the model, all of them are significant, suggesting that they all help explain expected return innovations. The PAPR estimates of the persistence of expected returns (γ_μ) show a more uniform pattern over the specifications than the corresponding OLS estimates (bottom row of each panel).

[Table 4 about here.]

This autoregressive coefficient of the expected return is a key parameter of the model, and obtaining results conditional on γ_μ is also informative, given the additional parameter uncertainty of PAPR compared to the standard predictive regression. Fixing the autoregressive parameter of the expected return process decreases the number of estimated parameters, reducing the parameter uncertainty in the PAPR. Results from restricted estimation are shown in Table 5. The first column replicates the unrestricted estimates, while in the second and third columns restricted estimation results are presented, imposing either $\gamma_\mu = 0.8$ or $\gamma_\mu = 0.95$. The different outcomes of the estimation show some variation, but the differences are not large. That is, the estimated innovation slope coefficients are not particularly sensitive to the restrictions. I return to the usefulness of imposing restrictions on γ_μ in the out-of-sample results discussed below.

[Table 5 about here.]

[10] In unreported results I have considered alternative specifications. In particular, I fitted higher-order autoregressive models for each predictor, where the order is determined by the Akaike and Bayesian information criteria, and ARMA(1,1) models. The results based on the alternative time-series specifications remain qualitatively similar.

To evaluate the PAPR, I also calculate its in-sample fit and real-time (out-of-sample) forecasting performance. The measures I use are the in-sample and out-of-sample R^2 defined in section 4, and the Diebold-Mariano test assessing equal forecasting performance (Diebold and Mariano, 1995). Table 6 presents both in-sample and out-of-sample results for the one regressor specifications and the full model, where all the regressors are included. Results for the standard linear model estimated by OLS, as well as the unrestricted PAPR and two of the restricted forms are shown. The full estimation of the system using maximum likelihood Kalman Filter is also presented.[11]

[Table 6 about here.]

Table 6 shows that the unrestricted PAPR outperforms the OLS estimation in each case in terms of in-sample R^2. This suggests that the predictive regression is misspecified and adjusting for persistence differences mitigates the misspecification. The in-sample gains of the PAPR range between 0.2 and 2 percentage points, the latter implying an 18 percent improvement on the standard predictive regression in terms of in-sample R^2. When restrictions are imposed on the PAPR, the in-sample results worsen somewhat compared to the unrestricted model.

As seen in the estimation results in Table 3, the standard errors of the PAPR estimates are relatively large, which might negatively affect the out-of-sample performance. This is at least partially supported by the out-of-sample results shown in Table 6. Imposing a predefined value on the persistence parameter γ_μ, as discussed above, can potentially reduce the overall parameter uncertainty and improve the out-of-sample forecasts. In particular, imposing a relatively high persistence ($\gamma_\mu = 0.95$) makes sense both from an economic and econometric perspective as most evidence suggest that time variation in expected returns is quite persistent. As seen in Table 6, fixing $\gamma_\mu = 0.95$, the PAPR forecasts perform the best out of sample in all cases except for the dividend-price ratio, which appears to be a weak predictor with no out-of-sample gains for any estimation method. In the multivariate specification, all PAPR forecasts (whether based on restricted or unrestricted estimates)

[11]Note that these results are not directly comparable with the results in Pástor and Stambaugh (2009), since performing the full Bayesian estimation as in the original paper is outside the scope of the current analysis.

outperform the OLS one. This is also in line with the fact that in the multivariate case the expected return parameters are estimated with more precision (see also in Panel (b) in Table 3).

The fit of the Kalman Filter tends to be much weaker than that of the other two methods in the specifications using the dividend-price ratio (Panel a and d in Table 6). This is likely due to the fact that the estimation of Kalman Filter parameters becomes unstable when the persistence of the state variables are high. Also, out-of-sample performance tends to be weak in all specifications, which is in line with Lucivjanska (2018) and the simulation results in Section 4.

6 Conclusion

In this paper, I investigate predictive regressions when the data are generated by the predictive system proposed by Pástor and Stambaugh (2009), where predictors are imperfect proxies of the expected returns. I show that predictor imperfection can be decomposed into two main terms, (i) the imperfect correlation between the innovation of the predictor and the expected return (ii) and the difference in persistence between the predictor and the expected returns. While the first type of imperfection is arguably "fundamental", the second type can be controlled for within the model. To this end, I propose a persistence adjustment to the standard predictive regression, which is based on the structural assumptions of the predictive system.

The proposed estimator is labeled Persistence Adjusted Predictive Regression (PAPR). It is a two-stage method, where the expected return and predictor processes are modelled separately, allowing for each to have distinct dynamic properties. This method represents a minimal deviation from the standard predictive regression. If the persistence parameters of the predictor and the expected return are equal, the method is asymptotically identical to the standard predictive regression. Simulations show that the model fit of the predictive regression can be substantially lower if the difference in persistence is not taken into account and the persistence adjustment can significantly improve upon standard least squares results in predictive regressions. This is particularly true if the difference in

persistence is large. Empirically, both in-sample and out-of-sample improvements, relative to OLS estimation, are documented in relevant cases.

The focus of the current paper is to evaluate how assumptions about imperfect predictors affect predictive regression evidence on return predictability. If the data-generating process is given by the predictive system, the Kalman Filter delivers asymptotically optimal expected return series. Disregarding estimation uncertainty, the PAPR is thus inferior to the Kalman Filter. However, a simple persistence adjustment brings the predictive regression results remarkably close to the estimation of the full system, and in practical situations the parameter uncertainty in the Kalman Filter results in poor in-sample and out-of-sample performance. The proposed method therefore provides a simple and almost efficient way of dealing with predictor imperfection.

Appendix

A t-statistics and R^2 of the predictive regression

Results similar to equation (7) can be derived for the t-statistic and the R^2 of the predictive regression. Start with the variance of the regression error:

$$\begin{aligned}
Var\left(\hat{e}_{t+1}\right) &= Var\left(r_{t+1}-\bar{r}\right)+\hat{\beta}_{OLS}^{2}Var\left(x_t-\bar{x}\right)-2\hat{\beta}_{OLS}Cov\left(r_{t+1}-\bar{r},x_t-\bar{x}\right)\\
&= \frac{\sigma_w^2}{1-\gamma_\mu^2}+\sigma_u^2+\hat{\beta}_{OLS}^{2}\frac{\sigma_\epsilon^2}{1-\gamma_x^2}-2\hat{\beta}_{OLS}\frac{\sigma_{w\epsilon}}{1-\gamma_x\gamma_\mu}\\
&= \frac{\sigma_w^2}{1-\gamma_\mu^2}+\sigma_u^2+\frac{\sigma_{w\epsilon}^2}{(1-\gamma_x\gamma_\mu)^2}\frac{(1-\gamma_x^2)^2}{\sigma_\epsilon^4}\frac{\sigma_\epsilon^2}{1-\gamma_x^2}-2\frac{\sigma_{w\epsilon}}{1-\gamma_x\gamma_\mu}\frac{(1-\gamma_x^2)}{\sigma_\epsilon^2}\frac{\sigma_{w\epsilon}}{1-\gamma_x\gamma_\mu}\\
&= \frac{\sigma_w^2}{1-\gamma_\mu^2}+\sigma_u^2-\frac{\sigma_{w\epsilon}^2}{(1-\gamma_x\gamma_\mu)^2}\frac{(1-\gamma_x^2)}{\sigma_\epsilon^2}.
\end{aligned}$$

The variance of the OLS estimator of β is

$$\begin{aligned}
TVar(\hat{\beta}_{OLS}) &= \frac{Var\left(\hat{e}_{t+1}\right)}{Var\left(x_t\right)}=\frac{\sigma_w^2}{\sigma_\epsilon^2}\frac{1-\gamma_x^2}{1-\gamma_\mu^2}+\frac{\sigma_u^2(1-\gamma_x^2)}{\sigma_\epsilon^2}-\rho_{w\epsilon}^2\frac{\sigma_w^2}{\sigma_\epsilon^2}\frac{(1-\gamma_x^2)^2}{(1-\gamma_x\gamma_\mu)^2}\\
&= \eta^2\frac{\sigma_u^2}{\sigma_\epsilon^2}(1-\gamma_x^2)+\frac{\sigma_u^2(1-\gamma_x^2)}{\sigma_\epsilon^2}-\eta^2\rho_{w\epsilon}^2\frac{\sigma_u^2}{\sigma_\epsilon^2}\frac{(1-\gamma_x^2)^2(1-\gamma_\mu^2)}{(1-\gamma_x\gamma_\mu)^2}\\
&= \frac{\sigma_u^2(1-\gamma_x^2)}{\sigma_\epsilon^2}\left[1+\eta^2\left(1-\rho_{w\epsilon}^2\frac{(1-\gamma_x^2)(1-\gamma_\mu^2)}{(1-\gamma_x\gamma_\mu)^2}\right)\right],
\end{aligned}$$

which implies the following expression for the t-statistic:

$$t_{\hat{\beta}_{OLS}}=\frac{\hat{\beta}_{OLS}}{\sqrt{Var(\hat{\beta}_{OLS})}}=\sqrt{T}\frac{\eta\rho_{w\epsilon}\frac{\sqrt{1-\gamma_\mu^2}\sqrt{1-\gamma_x^2}}{1-\gamma_x\gamma_\mu}}{\sqrt{1+\eta^2\left[1-\rho_{w\epsilon}^2\left(\frac{(1-\gamma_\mu^2)(1-\gamma_x^2)}{(1-\gamma_x\gamma_\mu)^2}\right)\right]}}.$$

The above expression is scale-free and essentially depends on the same factors as the slope coefficient: the underlying amount of predictability, the degree of imperfection and the difference between the persistence of the predictor and the expected return. Both η and $\rho_{w\epsilon}$ are positively related to the t-statistic. A higher underlying amount of predictability and a larger correlation between the predictor and the expected return

both suggest a stronger signal, which should imply a greater t-statistic. The effect of the autocorrelation parameters on the t-statistic is also similar. The t-statistic is maximal if the two autocorrelations coincide, otherwise it decreases in the difference between the two values. The decrease is faster the closer the parameters are to unity. It implies that the persistence of the regressor has a substantial effect on the inference, in line with the results in Appendix B.

A decomposition analogous to (7) can also be given for the regression R^2,

$$R^2_{OLS} = \frac{\hat{\beta}^2_{OLS} Var(x_t)}{Var(r_{t+1})} = R^2_{true} \rho^2_{w\epsilon} \frac{(1-\gamma_x^2)(1-\gamma_\mu^2)}{(1-\gamma_x\gamma_\mu)^2}.$$

$R^2_{OLS} \leq R^2_{true}$, so the explanatory power of *any* predictive regression provides a lower bound for how much expected returns explain return variation.

B Highly persistent predictors

Suppose that the autoregressive parameter of the predictor and the expected return is given by $\gamma_x = 1 - \frac{c_x}{T}$ and $\gamma_\mu = 1 - \frac{c_\mu}{T}$, where T is the sample size and c_x and c_μ are constants. This corresponds to the specification in Cavanagh et al. (1995) and Campbell and Yogo (2006). As discussed, predictor imperfection can stem from two sources, either a smaller-than-unity correlation in innovations, or a difference in the autocorrelation of the predictor and the expected returns (which in this case is captured by the fact that $c_x \neq c_\mu$).

Fixing the correlation coefficient $\rho_{w\epsilon}$ to a value strictly smaller than one implies a spurious predictive regression if the predictor and the expected returns are nearly integrated. In particular, any non-perfect correlation means that the expected return has a nearly integrated component that is unrelated to the predictor. In this case, the slope coefficient of the predictive regression converges to a random variable, and the t-statistic diverges for any value of $\rho_{w\epsilon}$. Therefore, it is necessary to introduce an asymptotically perfect correlation between the innovations to analyse the effect of predictor imperfection further. To this end I introduce the concept of a nearly perfect predictor (analogous to

near integration). It implies that asymptotically the correlation $\rho_{w\epsilon}$ converges to unity. In particular, I specify $\rho_{w\epsilon} = 1 - \frac{d}{T^\delta}$ for some positive constants d and δ. The asymptotic results on the regression slope coefficient and the t-statistics are summarized in the following proposition.

Proposition 1 *Assume that the data-generating process is given by equations* $(2)-(4)$*, with* $\gamma_x = 1 - \frac{c_x}{T}$ *and* $\gamma_\mu = 1 - \frac{c_\mu}{T}$ *for some* $c_x, c_\mu \geq 0$*, and the correlation between the predictor and the expected return can be written as* $\rho_{\epsilon w} = 1 - \frac{d}{T^\delta}$ *for some* $d > 0$ *and* $\delta \geq 0$*. Define* $\beta^* = \frac{\sigma_w}{\sigma_\epsilon}$*. The following results hold:*

1. *The slope coefficient satisfies*

$$
\begin{aligned}
T^{\delta/2}\left(\hat{\beta}_{OLS} - \beta^*\right) &\Rightarrow \sqrt{2d}\frac{\sigma_w \int \underline{J}_{c_x}\tilde{\underline{J}}_{c_\mu}}{\sigma_\epsilon \int \underline{J}_{c_x}^2} && \text{if } \delta < 2,\\
T\left(\hat{\beta}_{OLS} - \beta^*\right) &\Rightarrow \frac{\sigma_u \int \underline{J}_{c_x} dB_u}{\sigma_\epsilon \int \underline{J}_{c_x}^2} + \sqrt{2d}\frac{\sigma_w \int \underline{J}_{c_x}\tilde{\underline{J}}_{c_\mu}}{\sigma_\epsilon \int \underline{J}_{c_x}^2} && \text{if } \delta = 2,\\
T\left(\hat{\beta}_{OLS} - \beta^*\right) &\Rightarrow \frac{\sigma_u \int \underline{J}_{c_x} dB_u}{\sigma_\epsilon \int \underline{J}_{c_x}^2} && \text{if } \delta > 2.
\end{aligned}
$$

2. *The t-statistic*

$$
\begin{aligned}
&t_{\hat{\beta}_{OLS}} \text{ diverges} && \text{if } \delta < 2,\\
&t_{\hat{\beta}_{OLS}} \Rightarrow (1-\rho_{\epsilon u}^2)^{1/2}\frac{\int \underline{J}_{c_x} dW_{\tilde{u}}}{\left(\int \underline{J}_{c_x}^2\right)^{1/2}} + \rho_{\epsilon u}\frac{\int \underline{J}_{c_x} dW_x}{\left(\int \underline{J}_{c_x}^2\right)^{1/2}} + \sqrt{2d}\frac{\sigma_w \int \underline{J}_{c_x}\tilde{\underline{J}}_{c_\mu}}{\sigma_u\left(\int \underline{J}_{c_x}^2\right)^{1/2}} && \text{if } \delta = 2,\\
&t_{\hat{\beta}_{OLS}} \Rightarrow (1-\rho_{\epsilon u}^2)^{1/2}\frac{\int \underline{J}_{c_x} dW_{\tilde{u}}}{\left(\int \underline{J}_{c_x}^2\right)^{1/2}} + \rho_{\epsilon u}\frac{\int \underline{J}_{c_x} dW_x}{\left(\int \underline{J}_{c_x}^2\right)^{1/2}} && \text{if } \delta > 2,
\end{aligned}
$$

where the processes J_{c_x} *and* $\tilde{J}_{c_\mu}$ *are defined below in Lemma 1.*

Before the proof of Proposition 1, I collect some useful standard results related to local-to-unity asymptotics in a lemma (based on Phillips, 1987).

Lemma 1 *Let* $v_t = (u_t, w_t, \epsilon_t)'$ *be a serially uncorrelated martingale difference sequence with* $E(v_t v_t') = \Sigma$*. Define the nearly integrated processes* $\xi_t = (\tilde{u}_t, \mu_t, x_t)'$ *such that their*

time-series dynamics can be written as

$$\xi_t = \begin{pmatrix} 1 - \frac{c_{\tilde{u}}}{T} & 0 & 0 \\ 0 & 1 - \frac{c_\mu}{T} & 0 \\ 0 & 0 & 1 - \frac{c_x}{T} \end{pmatrix} \xi_{t-1} + v_t \equiv \left(I - \frac{1}{T} C \right) \xi_{t-1} + v_t.$$

Then by standard results (for example, in Phillips (1988))

$$\begin{aligned} \frac{1}{T^{-3/2}} \sum \xi_{t-1} &\Rightarrow \Sigma^{1/2} \int_0^1 \mathbf{J}_C(r) dr \\ \frac{1}{T^{-2}} \sum \xi_{t-1} \xi_{t-1}' &\Rightarrow \Sigma^{1/2} \int_0^1 \mathbf{J}_C(r) \mathbf{J}_C(r)' dr \Sigma^{1/2} \\ \frac{1}{T^{-1}} \sum \xi_{t-1} v_t' &\Rightarrow \Sigma^{1/2} \int_0^1 \mathbf{J}_C(r) d\mathbf{W}(r)' dr \Sigma^{1/2}, \end{aligned}$$

where $\mathbf{W}$ *is a standard three dimensional Brownian motion and* $\mathbf{J}_C$ *satisfies the stochastic differential equation*

$$d\mathbf{J}_C(r) = C\mathbf{J}_C(r) dr + d\mathbf{W}(r); \quad \mathbf{J}_C(0) = 0.$$

The joint convergence result also implies that the result holds for individual processes and pairs of processes, as well. Also note that due to the lack of autocorrelation in the innovations we can use the usual orthogonal decomposition for the processes. In particular, $W_\mu = \rho_{\epsilon w} W_x + \sqrt{1 - \rho_{\epsilon w}^2} W_{\tilde{\mu}}$, where $W_{\tilde{\mu}}$ is a standard Wiener process, independent from W_x, and using this decomposition for w_μ we can build up a $\tilde{J}_{c_\mu}$ process, which is independent from J_{c_x}. The results hold for the demeaned process $\underline{\xi}_t = \xi_t - \frac{1}{T} \sum_{t=1}^T \xi_t$ replacing $\mathbf{J}_C(r)$ with $\underline{\mathbf{J}}_C(r) = \mathbf{J}_C(r) - \int_0^1 \mathbf{J}_C(r) dr$.

Proof of Proposition 1.

Consider first the OLS estimate of the slope coefficient,

$$\begin{aligned}\hat{\beta}_{OLS} =& \frac{\sum (r_{t+1} - \bar{r})(x_t - \bar{x})}{\sum (x_t - \bar{x})^2} = \frac{\sum (\mu_t - \bar{\mu})(x_t - \bar{x})}{\sum (x_t - \bar{x})^2} + \frac{\sum u_{t+1}(x_t - \bar{x})}{\sum (x_t - \bar{x})^2} \\ =& \frac{\sum \left(1 - \frac{d}{T^\delta}\right) \frac{\sigma_w}{\sigma_\epsilon} (x_t - \bar{x})^2}{\sum (x_t - \bar{x})^2} + \frac{\sum u_{t+1}(x_t - \bar{x})}{\sum (x_t - \bar{x})^2} + \frac{\sum \sqrt{\frac{2d}{T^\delta} - \frac{d^2}{T^{2\delta}}} \left(\tilde{\mu}_{t+1} - \bar{\tilde{\mu}}\right)(x_t - \bar{x})}{\sum (x_t - \bar{x})^2} \\ =& \left(1 - \frac{d}{T^\delta}\right) \frac{\sigma_w}{\sigma_\epsilon} + \frac{\sum u_{t+1}(x_t - \bar{x})}{\sum (x_t - \bar{x})^2} + \frac{\sum \sqrt{\frac{2d}{T^\delta} - \frac{d^2}{T^{2\delta}}} \left(\tilde{\mu}_{t+1} - \bar{\tilde{\mu}}\right)(x_t - \bar{x})}{\sum (x_t - \bar{x})^2},\end{aligned}$$

where $\tilde{\mu}_{t+1}$ is a local-to-unity process with autoregressive parameter $1 - c_\mu/T$ and innovations uncorrelated with ϵ_t. By results in Lemma 1,

$$\begin{aligned} T^{\delta/2}\sqrt{\frac{2d}{T^\delta} - \frac{d^2}{T^{2\delta}}} \frac{\frac{1}{T}\sum \left(\tilde{\mu}_{t+1} - \bar{\tilde{\mu}}\right)(x_t - \bar{x})}{\frac{1}{T}\sum (x_t - \bar{x})^2} &\Rightarrow \sqrt{2d}\frac{\sigma_w \int \underline{J}_{c_x} \underline{\tilde{J}}_{c_\mu}}{\sigma_\epsilon \int \underline{J}_{c_x}^2}, \\ T\frac{\frac{1}{T}\sum u_{t+1}(x_t - \bar{x})}{\frac{1}{T}\sum (x_t - \bar{x})^2} &\Rightarrow \frac{\sigma_u \int \underline{J}_{c_x} dW_u}{\sigma_\epsilon \int \underline{J}_{c_x}^2}, \end{aligned}$$

and joint weak convergence also holds. Defining $\beta^* = \frac{\sigma_w}{\sigma_\epsilon}$ gives the first part of the results in Proposition 1. The result implies that the OLS estimate is consistent,

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta,$$

and $\hat{\beta}_{OLS} - \beta = O_p(T^{-\min\{1,\delta/2\}})$, which can also be written as $\hat{\beta}_{OLS} - \beta = \frac{\zeta}{T^{\min\{1,\delta/2\}}}$, where ζ is a random variable ($O_p(1)$). The fitted residuals and their sum of squares can thus be written as,

$$\begin{aligned}\hat{u}_{t+1} &= r_{t+1} - \hat{\beta}_{OLS}(x_t - \bar{x}) \\ &= \mu_t - \bar{\mu}_t + u_{t+1} - \bar{u}_t - \hat{\beta}_{OLS}(x_t - \bar{x}) \\ &= \left(1 - \frac{d}{T^\delta}\right)\beta(x_t - \bar{x}) + \sqrt{\frac{2d}{T^\delta} - \frac{d^2}{T^{2\delta}}}(\tilde{\mu} - \bar{\tilde{\mu}}) - \hat{\beta}_{OLS}(x_t - \bar{x}) + (u_{t+1} - \bar{u}) \\ &= \left(\beta - \hat{\beta}_{OLS} - \frac{d}{T^\delta}\right)(x_t - \bar{x}) + \sqrt{\frac{2d}{T^\delta} - \frac{d^2}{T^{2\delta}}}(\tilde{\mu} - \bar{\tilde{\mu}}) + (u_{t+1} - \bar{u})\end{aligned}$$

and,

$$\begin{aligned}\frac{1}{T}\sum \hat{u}_{t+1}^2 &= \frac{1}{T}\left(\beta-\hat{\beta}_{OLS}-\frac{d}{T^\delta}\right)^2\sum(x_t-\bar{x})^2+\frac{1}{T}\left(\frac{2d}{T^\delta}-\frac{d^2}{T^{2\delta}}\right)\sum\left(\tilde{\mu}-\bar{\tilde{\mu}}\right)^2+\frac{1}{T}\sum(u_{t+1}-\bar{u})^2\\ &+2\frac{1}{T}\left(\beta-\hat{\beta}_{OLS}-\frac{d}{T^\delta}\right)\sqrt{\frac{2d}{T^\delta}-\frac{d^2}{T^{2\delta}}}\sum\left(x_t-\bar{x}\right)\left(\tilde{\mu}-\bar{\tilde{\mu}}\right)\\ &+2\frac{1}{T}\left(\beta-\hat{\beta}_{OLS}-\frac{d}{T^\delta}\right)\sum(x_t-\bar{x})(u_{t+1}-\bar{u})\\ &+2\frac{1}{T}\sqrt{\frac{2d}{T^\delta}-\frac{d^2}{T^{2\delta}}}\sum\left(\tilde{\mu}-\bar{\tilde{\mu}}\right)\left(u_{t+1}-\bar{u}\right)\\ &=\frac{1}{T}\sum(u_{t+1}-\bar{u})^2+o_p(1).\end{aligned}$$

The variance of the OLS estimator of β can now be written as

$$Var(\hat{\beta}_{OLS})=\frac{1}{T}\frac{\frac{1}{T}\sum\hat{u}_{t+1}^2}{\frac{1}{T}\sum(x_t-\bar{x})^2}=\frac{1}{T}\frac{\zeta^2\left(o_p(1)+\frac{1}{T\zeta^2}\sum(u_{t+1}-\bar{u})^2\right)}{\frac{1}{T}\sum(x_t-\bar{x})^2}$$

and the t-statistic (disregarding the asymptotically vanishing term) is

$$t_{\hat{\beta}_{OLS}}=\frac{\hat{\beta}_{OLS}-\beta}{\sqrt{Var(\hat{\beta}_{OLS})}}=\frac{\frac{\zeta}{T^{\min\{1,\delta/2\}}}}{\frac{\zeta}{\sqrt{T}}\sqrt{\frac{\frac{1}{T\zeta^2}\sum(u_{t+1}-\bar{u})^2}{\frac{1}{T}\sum(x_t-\bar{x})^2}}}=T^{\max\{-\frac{1}{2},\frac{1-\delta}{2}\}}\frac{\zeta\sqrt{\frac{1}{T}\sum(x_t-\bar{x})^2}}{\sqrt{\frac{1}{T}\sum(u_{t+1}-\bar{u})^2}}.$$

If $\delta<2$, the above expression diverges, establishing the first line of the second part of the results. Otherwise, the above expression converges to

$$t_{\hat{\beta}_{OLS}}\Rightarrow\frac{\zeta\left(\sigma_\epsilon^2\int\underline{J}_{c_x}^2\right)^{1/2}}{\sigma_u}=\frac{\int\underline{J}_{c_x}dW_u}{\left(\int\underline{J}_{c_x}^2\right)^{1/2}}+\sqrt{2d}\frac{\sigma_w\int\underline{J}_{c_x}\tilde{\underline{J}}_{c_\mu}}{\sigma_u\left(\int\underline{J}_{c_x}^2\right)^{1/2}}\qquad\text{if }\delta=2,$$

and

$$t_{\hat{\beta}_{OLS}}\Rightarrow\frac{\zeta\left(\sigma_\epsilon^2\int\underline{J}_{c_x}^2\right)^{1/2}}{\sigma_u}=\frac{\int\underline{J}_{c_x}dW_u}{\left(\int\underline{J}_{c_x}^2\right)^{1/2}}\qquad\text{if }\delta>2.$$

The remaining results are then established by decomposing the innovations in the last term on the right hand side as $W_u=\rho_{eu}W_x+(1-\rho_{eu}^2)^{1/2}W_{\tilde{u}}$, where $W_{\tilde{u}}$ is a standard Wiener process independent of W_x. ■

Proposition 1 distinguishes three cases. First, if the predictor imperfection is strong ($\delta < 2$), then spurious regression dominates. Here the convergence of β is slow (or it does not converge in probability if $\delta = 0$), and the t-statistic diverges, showing strong statistical significance, even if there is no relationship between the predictor and the expected return. That is, the predictive regression is spurious if $\delta < 2$. It is analogous to the result in Ferson et al. (2003) and Deng (2013) discussing completely spurious regressions in the context of return predictability ($\delta = 0$ and $d = 1$ nests the specification discussed in Deng, 2013). In this case, the endogeneity of the predictor has a second order effect. The second case, $\delta > 2$, turns out to be asymptotically equivalent to the situation where the predictor is perfectly correlated with the expected return. Estimation of β is consistent and the t-statistic converges to the expression derived in Cavanagh et al. (1995). Third, in the knife-edge case, $\delta = 2$, predictor imperfection has a non-trivial effect on the results. The slope coefficient can be estimated consistently, but the t-statistic becomes non-standard. In particular, it is a linear combination of three terms. The first one is a standard normal distribution reflecting that in the absence of endogeneity and imperfection ($\rho_{\epsilon w} = d = 0$), the t-statistic asymptotically follows a standard normal distribution. The second part enters because of the potential endogeneity of the regressor and the third component indicates the effect of predictor imperfection. The relative strength of the three terms are determined by the correlation structure of the innovations.

C The general formulation of the two-step procedure

Let the predictive system be described by

$$r_{t+1} = \mu_t + u_{t+1} \tag{13}$$

$$\mu_{t+1} = (1 - \gamma_\mu)\bar{\mu} + \gamma_\mu \mu_t + w_{t+1} \tag{14}$$

$$\mathbf{x}_{t+1} = \bar{\mathbf{x}} + A(L)\mathbf{x}_t + \boldsymbol{\epsilon}_\mathbf{t}, \tag{15}$$

where $\mathbf{x}_t$ is a K-dimensional vector process, and $A(L)$ is a general lag polynomial satisfying assumptions that ensure the stationarity of $\mathbf{x}_t$. The innovations $\{u_t, w_t, \boldsymbol{\epsilon}_\mathbf{t}\}$ jointly follow

martingale difference sequences with contemporaneous covariance matrix Σ. Let $\theta = (\theta_1, \theta_2)$ be the parameters of the model to be estimated, with $\theta_1 = \{\bar{\mathbf{x}}, A\}$ and $\theta_2 = \{\bar{\mu}, \gamma_\mu, \mathbf{b}\}$ as in the main text.

In the first step, the parameters of equation (15) are estimated (for example equation-by-equation OLS), and the predictor innovations are calculated as

$$\hat{\boldsymbol{\epsilon}}_t = \mathbf{x}_{t+1} - \hat{\bar{\mathbf{x}}} - \hat{A}(L)\mathbf{x}_t$$

The second step consists of minimization of the criterion function

$$\hat{\theta}_2 = \arg\min_{\theta_2} Q(\hat{\theta}_1, \theta_2) = \arg\min_{\theta_2} \frac{1}{T-1} \sum_{t=1}^{T-1} \left(r_{t+1} - \bar{\mu} + \sum_{s=1}^{t} \gamma_\mu^{t-s} \mathbf{b}' \hat{\boldsymbol{\epsilon}}_t \right)^2, \qquad (16)$$

The asymptotic properties of the two-step estimator, which is a special case of extremum estimators, is well-known (for example Newey and McFadden, 1994). In particular, given a consistent estimator for the first step parameters $\hat{\theta}_1$, the remaining parameters (θ_2) can be consistently estimated in the second step, treating the first step parameters as known.

The asymptotic variance of the second step estimator is, however, affected by the first step results. In particular, the standard errors of the second step estimates have to account for the fact that the first step parameters are estimated. Let $y_t = \{r_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}, ...\}$ denote the data, and

$$q(\theta_1, \theta_2, y_t) = \left(r_{t+1} - \bar{\mu} + \sum_{s=1}^{t} \gamma_\mu^{t-s} \mathbf{b}' \boldsymbol{\epsilon}_\mathbf{t} \right)^2,$$

where the dependence on θ_1 is through the innovations $\boldsymbol{\epsilon}_\mathbf{t}$. Also, let the random variable $r_t(\theta_1)$ be defined as

$$\sqrt{T}(\hat{\theta}_1 - \theta_1) = T^{-1/2} \sum_{t=1}^{T} r_t(\theta_1) + \mathrm{o}_p(1).$$

Then the asymptotic distribution of the second step estimator is given by

$$\begin{aligned}
\sqrt{T}(\hat{\theta}_2 - \theta_2) &\xrightarrow{d} N(0, \Gamma_0^{-1}\Omega_0\Gamma_0^{-1}), \\
\Gamma_0 &= E\left(\frac{\partial^2 q(\theta_1, \theta_2, y_t)}{\partial\theta_2\partial\theta_2'}\right), \\
\Omega_0 &= E\left(s(\theta_1, \theta_2, y_t)s(\theta_1, \theta_2, y_t)'\right), \\
s(\theta_1, \theta_2, y_t) &= \frac{\partial q(\theta_1, \theta_2, y_t)}{\partial\theta_2} + \frac{\partial^2 q(\theta_1, \theta_2, y_t)}{\partial\theta_2\partial\theta_1'} r_t(\theta_1).
\end{aligned}$$

The effect of the first step estimation appears through the second term of the $s(\theta_1, \theta_2, y_t)$ function (in a usual one-step estimator $\frac{\partial^2 q(\theta_1, \theta_2, y_t)}{\partial\theta_2\partial\theta_1'} = 0$). When calculating the standard errors, the matrices Γ_0, Ω_0 are replaced by consistent estimators

$$\begin{aligned}
\hat{\Gamma}_0 &= \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 q(\hat{\theta}_1, \theta_2, y_t)}{\partial\theta_2\partial\theta_2'}|_{\theta_2=\hat{\theta}_2}, \\
\hat{\Omega}_0 &= \frac{1}{T}\sum_{t=1}^{T} \hat{s}(\hat{\theta}_1, \hat{\theta}_2, y_t)\hat{s}(\hat{\theta}_1, \hat{\theta}_2, y_t)', \\
\hat{s}(\hat{\theta}_1, \hat{\theta}_2, y_t) &= \frac{\partial q(\hat{\theta}_1, \theta_2, y_t)}{\partial\theta_2}|_{\theta_2=\hat{\theta}_2} + \frac{\partial^2 q(\theta_1, \theta_2, y_t)}{\partial\theta_2\partial\theta_1'}|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} r_t(\hat{\theta}_1).
\end{aligned}$$

References

Ang, A. and Bekaert, G. (2007). Stock return predictability: Is it there? *Review of Financial Studies*, 20(3):651–707.

Campbell, J. Y. and Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, 1(3):195–228.

Campbell, J. Y. and Yogo, M. (2006). Efficient tests of stock return predictability. *Journal of Financial Economics*, 81(1):27–60.

Cavanagh, C. L., Elliott, G., and Stock, J. H. (1995). Inference in models with nearly integrated regressors. *Econometric Theory*, 11(05):1131–1147.

Cochrane, J. H. (2008). The dog that did not bark: A defense of return predictability. *Review of Financial Studies*, 21(4):1533–1575.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66(4):1047–1108.

Deng, A. (2013). Understanding Spurious Regression in Financial Economics. *Journal of Financial Econometrics*, 12(1):122–150.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):134–144.

Ferson, W. E., Sarkissian, S., and Simin, T. T. (2003). Spurious Regressions in Financial Economics? *Journal of Finance*, 58(4):1393–1414.

Goyal, A. and Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508.

Jansson, M. and Moreira, M. J. (2006). Optimal inference in regression models with nearly integrated regressors. *Econometrica*, 74(3):681–714.

Koijen, R. S. and Van Nieuwerburgh, S. (2011). Predictability of returns and cash flows. *Annual Review of Financial Economics*, 3:467–491.

Kostakis, A., Magdalinos, T., and Stamatogiannis, M. P. (2015). Robust econometric inference for stock return predictability. *Review of Financial Studies*, 28(5):1506–1553.

Lettau, M. and Ludvigson, S. (2001). Consumption, aggregate wealth, and expected stock returns. *the Journal of Finance*, 56(3):815–849.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics*, 74(2):209–235.

Lucivjanska, K. (2018). Is imperfection better? evidence from predicting stock and bond returns. *Journal of Financial Econometrics*, 16(2):244–270.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Pástor, L. and Stambaugh, R. F. (2009). Predictive systems: Living with imperfect predictors. *The Journal of Finance*, 64(4):1583–1628.

Phillips, P. C. (1988). Regression theory for near-integrated time series. *Econometrica: Journal of the Econometric Society*, pages 1021–1043.

Phillips, P. C. (2014). On confidence intervals for autoregressive roots and predictive regression. *Econometrica*, 82(3):1177–1195.

Phillips, P. C. B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika*, 74(3):535–547.

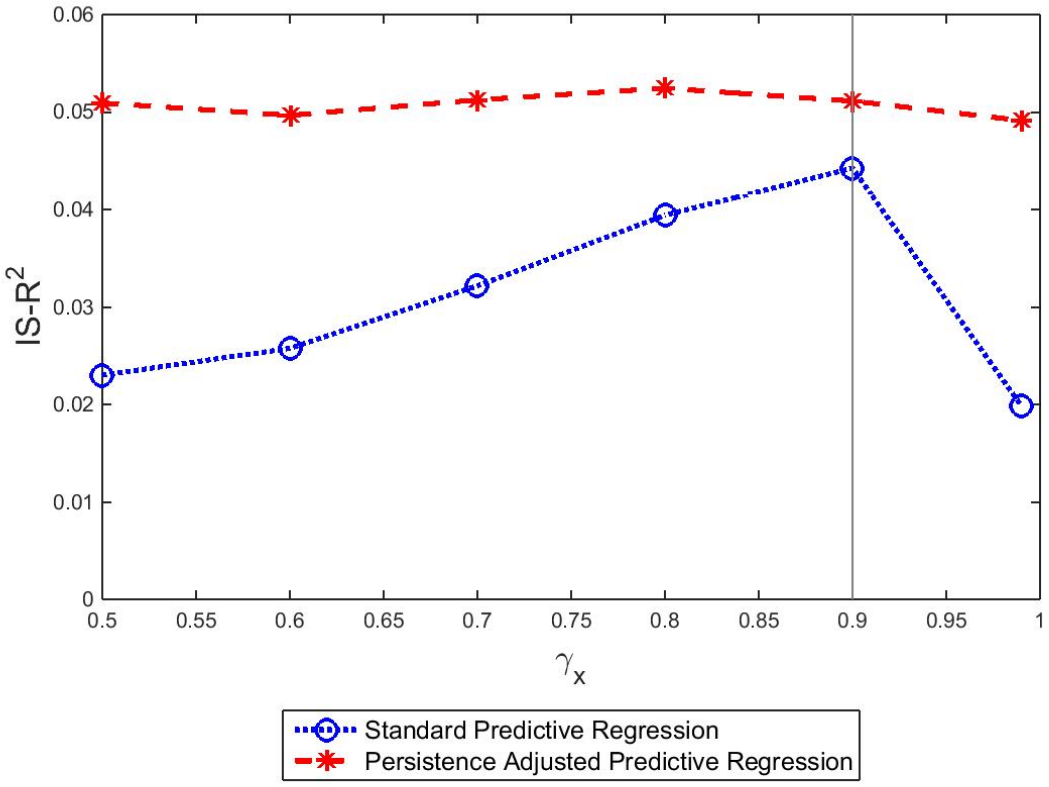
Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, 54(3):375–421.

Torous, W., Valkanov, R., and Yan, S. (2004). On predicting stock returns with nearly integrated explanatory variables. *The Journal of Business*, 77(4):937–966.

Wachter, J. A. and Warusawitharana, M. (2009). Predictable returns and asset allocation: Should a skeptical investor time the market? *Journal of Econometrics*, 148(2):162–178.

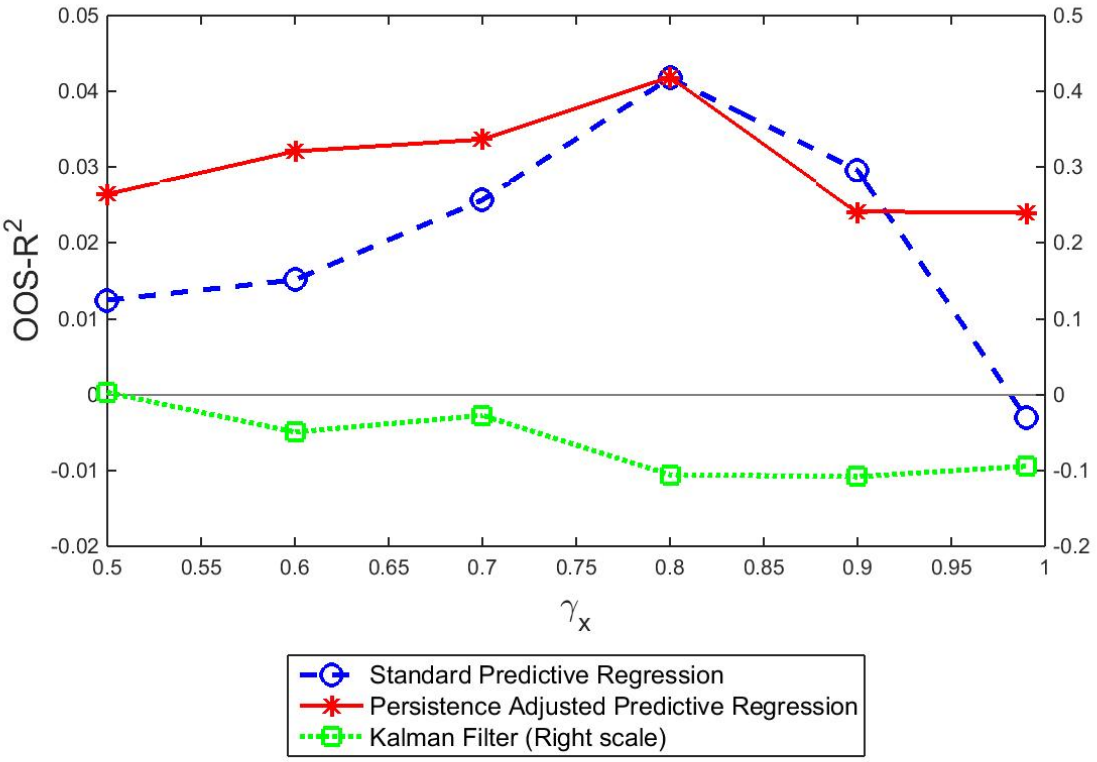
Wachter, J. A. and Warusawitharana, M. (2015). What is the chance that the equity premium varies over time? evidence from regressions on the dividend-price ratio. *Journal of Econometrics*, 186(1):74–93.

Figure 2: **Simulated in-sample model fit**



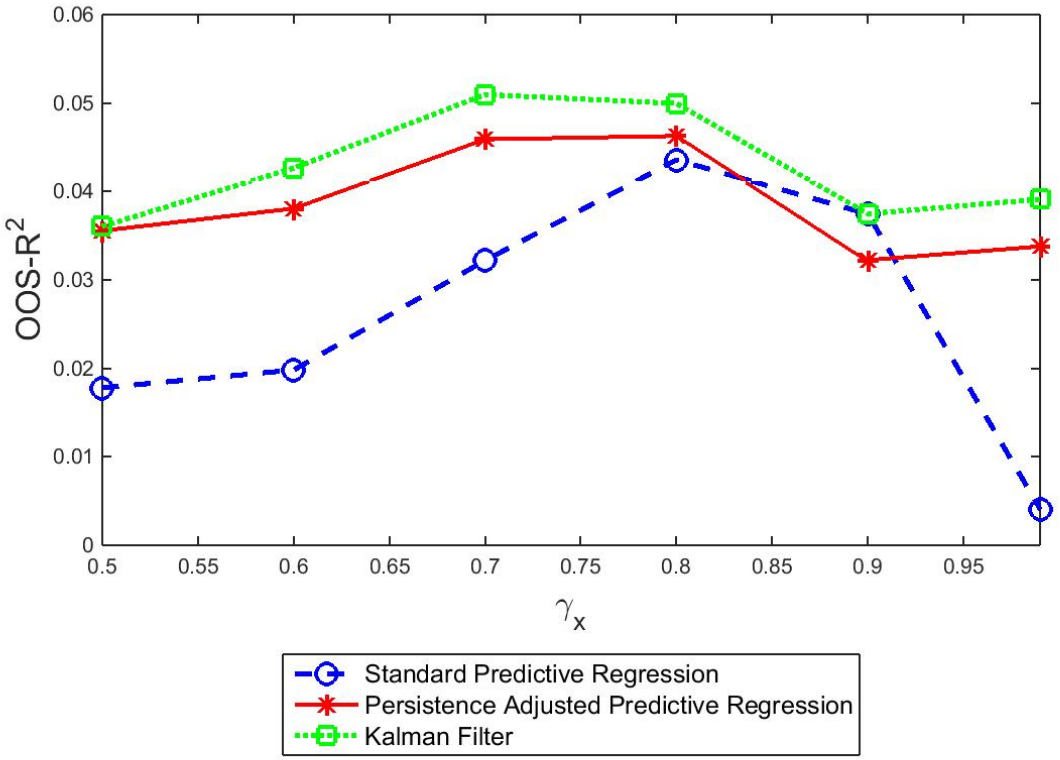
Notes: This plot shows the in-sample R^2 of the predictive regression and the Persistence Adjusted Predictive Regression as a function of the autoregressive parameter of the predictor. The results are based on a Monte Carlo simulation with 1000 repetitions. The parameter choices are as in the description for Table 1. In particular, the vertical line indicates that the autoregressive parameter of the expected return is set to $\gamma_\mu = 0.9$.

Figure 3: **Simulated Out-of-sample model fit when parameters are estimated**



Notes: These plots show the out-of-sample R^2 of the standard predictive regression (dashed line), the Persistence Adjusted Predictive Regression (solid line) and the Kalman Filter (dotted line). The autoregressive parameter of the expected return is set to $\gamma_\mu = 0.9$, and results are shown as a function of the persistence parameter of the predictor. The other parameters are set to their default values as described in the text and in Table 1. The benchmark model is the historical mean forecast and the results are based on 1000 repetitions.

Figure 4: **Simulated out-of-sample model fit when parameters are imposed**



Notes: This plot shows the out-of-sample R^2 of the standard predictive regression (dashed line), the Persistence Adjusted Predictive Regression (solid line) and the Kalman Filter (dotted line), where the parameters are not estimated (the true parameters are imposed). The autoregressive parameter of the expected return is set to $\gamma_\mu = 0.9$, and results are shown as a function of the persistence parameter of the predictor. The other parameters are set to their default values as described in the text and in Table 1. The benchmark model is the historical mean forecast and the results are based on 1000 repetitions.

Table 1: **Baseline simulation**

This table shows simulation results for the baseline model. Panel (a) contains results based on the standard predictive regression estimated by OLS, and Panel (b) shows results for the Persistence-Adjusted Predictive Regression. In Panel (a) (Panel b) the first two columns are the mean and standard deviation of the (implied) predictive regression coefficient estimates. The implied predictive regression coefficient is calculated as $\hat{\beta}_{Implied} = \hat{b}\frac{1-\hat{\gamma}_x^2}{1-\hat{\gamma}_\mu\hat{\gamma}_x}$, where $\hat{b}$, $\hat{\gamma}_x$ and $\hat{\gamma}_\mu$ are estimates of the Persistence Adjusted Predictive Regression. The third and fourth columns are the mean and standard deviation of the autoregressive coefficient of the expected return, measured by the sample first order autocorrelation. The data-generating process is the predictive system in equation (2)-(4). Each row indicates the persistence parameter of the predictor used in the simulation. Otherwise, default parameter values are used: $\bar{\mu} = 0.018$, $\bar{x} = 0.03$, $\eta^2 = 0.05$, $\sigma_u = 0.081$, $\gamma_\mu = 0.9$, $\sigma_w = 0.011$,$\rho_{uw} = -0.7$,$\rho_{w\epsilon} = 0.9$,$\sigma_\epsilon = 0.011\frac{1-\gamma_x^2}{1-0.9\gamma_x}$. All the results are based on a sample size of $T = 200$ and 1000 repetitions.

(a) Predictive Regression

	Mean $\hat{\beta}$	**S.e.** $\hat{\beta}$	**Mean** $\hat{\gamma}_\mu$	**S.e.** $\hat{\gamma}_\mu$
$\gamma_x = 0.5$	0.9362	0.4468	0.4862	0.0622
$\gamma_x = 0.6$	0.9156	0.4006	0.5843	0.0595
$\gamma_x = 0.7$	0.9403	0.3679	0.6846	0.0519
$\gamma_x = 0.8$	0.9720	0.3430	0.7817	0.0448
$\gamma_x = 0.9$	0.9999	0.3328	0.8810	0.0343
$\gamma_x = 0.99$	1.3737	1.0170	0.9772	0.0177

(b) Persistence Adjusted Predictive Regression

	Mean $\hat{\beta}$	**S.e.** $\hat{\beta}$	**Mean** $\hat{\gamma}_\mu$	**S.e.** $\hat{\gamma}_\mu$
$\gamma_x = 0.5$	1.0680	0.4125	0.8051	0.2266
$\gamma_x = 0.6$	1.0288	0.3684	0.8224	0.2092
$\gamma_x = 0.7$	1.0518	0.3514	0.8395	0.1681
$\gamma_x = 0.8$	1.0701	0.3426	0.8495	0.1497
$\gamma_x = 0.9$	1.0936	0.3675	0.8571	0.1476
$\gamma_x = 0.99$	1.6635	1.3448	0.8083	0.1788

Table 2: **Descriptive statistics**

This table includes descriptive statistics for variables used in the main empirical analysis. The data is quarterly, running between the first quarter of 1952 and the fourth quarter of 2016. The first two columns are the mean and the standard deviation of the variables. The third column is the estimated slope coefficient of a first order autoregressive process.

	mean	**stdev**	γ_x	**N**
dp	0.0308	0.0111	0.967	260
cay	-2.07e-05	0.0125	0.822	260
by	7.12e-05	0.00531	0.612	260
ret	0.0181	0.0824	0.0821	260

Table 3: **Estimation results**

This table shows estimation results for the predictive regression with and without persistence adjustment. The first three columns show results based on one predictor while the last column shows the results when all predictors are included. Panel (a) includes the estimates of univariate (columns 1-3) and multivariate OLS regressions. The slope coefficient estimates are shown in rows one through three. The last row shows the implied autocorrelation of the expected return, that is the first order autocorrelation of the process $\hat{\mu}_t = \sum_{j=1}^{J} \hat{\beta}_j x_j$. Standard errors are given in parentheses. Panel (b) shows results for the Persistence Adjusted Predictive Regressions. The first three rows are the innovation slope coefficients, and the last row is the estimated persistence of the expected return. Innovations in the first step are obtained through a first order autoregressive filter. The estimation method is non-linear least squares, and standard errors in parentheses are calculated using the asymptotic formula given in Appendix C. The sample runs between the first quarter of 1952 and the last quarter of 2016. The dependent variable is the one-step ahead excess return. *, ** and *** mean significance on a 10, 5 and 1 percent level, respectively.

(a) Predictive Regression

	dp	**cay**	**by**	**full**
β_{dp}	0.9196			0.7655
	(0.4649)**			(0.4604)*
β_{cay}		1.6051		1.2921
		(0.4106)***		(0.4230)***
β_{by}			2.9109	2.7117
			(1.1792)**	(1.1242)**
γ_{μ}	0.9644	0.8248	0.6145	0.7768
	(0.0147)***	(0.0340)***	(0.0849)***	(0.0423)***

(b) Persistence-Adjusted Predictive Regression (Second step)

	dp	**cay**	**by**	**full**
b_{dp}	2.1933			3.0026
	(1.5889)			(1.0243)***
b_{cay}		1.2293		0.9136
		(0.3658)***		(0.4169)**
b_{by}			2.7669	2.0009
			(1.2236)**	(0.8130)**
γ_{μ}	0.8762	0.9235	0.6808	0.9165
	(0.1236)***	(0.0395)***	(0.2155)***	(0.0444)***

Table 4: **Persistence Adjusted Predictive Regression (first stage)**

This table shows the first step estimates of the two-step procedures described in the text. The sample runs between the first quarter of 1952 and the last quarter of 2016. Since a first order autoregressive process is used, the estimated parameters are the intercept (second column) and the scalar autoregressive parameter (third column). Standard errors are based on the usual OLS formula. *, ** and *** mean significance on a 10, 5 and 1 percent level, respectively.

	Constant	**AR(1)**
dp	0.0009	0.9671
	(0.0007)	(0.0169)*
cay	-0.0000	0.8223
	(0.0004)	(0.0376)*
by	0.0000	0.6121
	(0.0003)	(0.0293)*

Table 5: **Persistence Adjusted Predictive Regression (restricted estimates)**

This table shows how restrictions affect estimation results for the two-step estimates. The sample runs between the first quarter of 1952 and the last quarter of 2016. The first column is the unrestricted model, while results in the second and third columns are obtained by fixing the autoregressive coefficient of the expected return. Innovations in the first step are obtained through a first order autoregressive filter. The dependent variable is the one-step ahead excess return. The estimation method is non-linear least squares, and standard errors in parenthesis are calculated using the asymptotic formula given in Appendix C. *, ** and *** mean significance on a 10, 5 and 1 percent level, respectively.

	Full	**Restricted** ($\gamma_\mu = 0.8$)	**Restricted** ($\gamma_\mu = 0.95$)
dp	3.0026	2.7873	2.6010
	(1.0243)***	(1.4966)*	(0.8032)***
cay	0.9136	1.1948	0.7950
	(0.4169)**	(0.6610)*	(0.3419)**
by	2.0009	2.5758	1.5699
	(0.8130)**	(1.0418)**	(0.7087)**
γ_μ	0.9165	0.8000	0.9500
	(0.0444)***		

Table 6: **Model comparison**

These tables presents performance measures for different model specifications. Panel (a) to (c) include results based on univariate specifications using the dividend-price ratio, the cay and the bond yield, respectively. Panel (d) shows results when all the variables are included simultaneously. The first column in each table presents in-sample R^2 values based on the formula in equation (11). The second column is the out-of-sample R^2 defined in equation (12). Positive values indicate that the model performs better than the historical mean. The first row indicates standard predictive regression results, the second row is the unrestricted Persistence Adjusted Predictive Regression (PAPR). The third and fourth rows are restricted versions of the PAPR, where the persistence of the expected return is fixed. The last row shows results based on the maximum likelihood estimates of the full system, using Kalman Filter. Only one-step ahead forecast horizons are considered, and the initial training sample is taken to be the 40 percent of the entire sample. The third column is the test statistic of the Diebold-Mariano test as in Diebold and Mariano (1995).

(a) dividend-price ratio

	IS-R^2	**OOS-R^2**	**DM test statistic**
OLS	0.0153	-0.0299	-7.7337
PAPR	0.0185	-0.0486	-12.8341
Restricted PAPR ($\gamma_\mu = 0.8$)	0.0150	-0.0069	-2.5791
Restricted PAPR ($\gamma_\mu = 0.95$)	0.0140	-0.0355	-9.9231
Kalman Filter	0.0069	-0.7463	-14.2686

(b) cay

	IS-R^2	**OOS-R^2**	**DM test statistic**
OLS	0.0587	0.0401	7.3666
PAPR	0.0685	0.0367	6.9956
Restricted PAPR ($\gamma_\mu = 0.8$)	0.0559	0.0307	5.6796
Restricted PAPR ($\gamma_\mu = 0.95$)	0.0679	0.0504	10.6227
Kalman Filter	0.0870	0.0037	0.6103

(c) bond yield

	IS-R^2	**OOS-R^2**	**DM test statistic**
OLS	0.0425	-0.0050	-0.7302
PAPR	0.0449	-0.0210	-3.0038
Restricted PAPR ($\gamma_\mu = 0.8$)	0.0444	-0.0054	-0.8058
Restricted PAPR ($\gamma_\mu = 0.95$)	0.0321	0.0047	0.9358
Kalman Filter	0.0442	-0.0806	-9.1279

(d) all

	IS-R^2	**OOS-R^2**	**DM test statistic**
OLS	0.0927	-0.0049	-0.6252
PAPR	0.1114	0.0087	1.0897
Restricted PAPR ($\gamma_\mu = 0.8$)	0.0944	0.0069	0.8341
Restricted PAPR ($\gamma_\mu = 0.95$)	0.1093	0.0217	3.0534
Kalman Filter	0.0186	-0.8611	-15.2522