Predictive Regression With Order-p Autoregressive Predictor

Yakov Amihud [*] Clifford M. Hurvich[†] Yi Wang[‡]

November 25, 2009

Abstract

Studies of predictive regressions analyze the case where y_t is predicted by x_{t-1} with x_t being first-order autoregressive, AR(1). Under some conditions, the OLS-estimated predictive coefficient is known to be biased. We analyze a predictive model where y_t is predicted by $x_{t-1}, x_{t-2}, \ldots x_{t-p}$ with x_t being autoregressive of order p, AR(p) with $p > 1$. We develop a generalized augmented regression method that produces a reduced-bias point estimate of the predictive coefficients and derive an appropriate hypothesis testing procedure. We apply our method to the prediction of quarterly stock returns by dividend yield, which is apparently AR(2). Using our method results in the AR(2) predictor series having insignificant effect, although under OLS, or the commonly assumed AR(1) structure, the predictive model is significant. We also generalize our method to the case of multiple $AR(p)$ predictors.

JEL classification: C32; G12

Keywords: Autoregressive, Augmented Regression Method (ARM)

[*]Ira Leon Rennert Professor of Finance, Stern School of Business, New York University, New York NY 10012; Tel: 212-998-0720; Fax: 212-995-4220; Email address: yamihud@stern.nyu.edu.

[†]Corresponding author. Professor of Statistics and Research Professor of Information, Operations and Management Science, Stern School of Business, New York University, New York NY 10012

[‡]Ph.D. Student, Operations and Management Science, Stern School of Business, New York University, New York NY 10012

I Introduction

Consider a predictive regression model where y_t is regressed on a lagged predictor variable, x_{t-1}. The OLS-estimated slope coefficient in such a model has been shown to be biased in small samples when the predictor x_t is first-order autoregressive (AR(1)) and the errors of the autoregressive model for x_t are correlated with the errors in the predictive regression model. Stambaugh(1999) analyzes this case and develops the bias expression.[1]

Research on this topic commonly assumes that the predictor series is AR(1). This is indeed the evidence on monthly series of some popular predictor series, such as dividend yield, earnings/price ratio and book/market ratio. However, other predictor series may be autoregressive of higher order. Even the dividend yield series is found in our study below to be AR(2) when examined at quarterly frequency. Here, we present a methodology which focuses on estimating and testing the predictive coefficients of a predictor which is AR(p), $p \geq 1$.

We analyze the case where y_t is predicted by $x_{t-1}, ..., x_{t-p}$ where the series x_t is AR(p), $p \geq 1$. We propose a reduced-bias method of estimating slope coefficients and a corresponding hypothesis test. This allows for testing predictive models with more general dynamic structure than those previously studied, which solve the problem only for the AR(1) case. The method developed here is a generalization of the Augmented Regression Method of Amihud and Hurvich (2004) for the AR(1) case.[2]

Predictive regressions with autoregressive predictors that are not necessarily of the AR(1) structure are quite common in finance and economics. Ferson, Sarkissian and Simon (2003, Table 1) provide a list of variables that are commonly used to predict

[1]See also Mankiw and Shapiro (1986), Nelson and Kim (1993).

[2]Amihud, Hurvich and Wang (2008) propose a hypothesis testing method for *multi*-predictor regression where the predictor vector x_t is assumed to following a vector AR(1), VAR(1), structure.

asset returns, some of which may not follow AR(1) structure. These include Fama and French's (1989) model where the cumulative stock return over various time intervals is predicted by dividend yield, the term bond-yield spread or the default bond-yield spread, all being autoregressive. Fama (1990) uses a closely related model. We show below the use of our method for quarterly dividend yield as a predictor variable, which apparently is AR(2). Patelis (1997) studies the effect of monetary policy on stock returns by regressing stock returns over various time intervals—1, 3, 12 or 24 months—on lagged values of the federal funds rate, various term yield spreads and one-month real interest rate. Ang and Bekaert (2007), using international data, find that dividend yield predicts stock returns at short horizons only when augmented with the short-term interest rate. Ferson et al. (2003) point out that while stock returns are not highly persistent, *expected* return may be persistent and thus may be spuriously predicted by an autoregressive series.

In economics, predictive regression studies are quite prevalent, with some predictor variables having autoregressive structure of order that is apparently greater than 1. Some models predict real activity, such as GNP (or GDP) growth, by term yield spread (Estrella and Hardouvelis (1991) and studies that followed). Plosser and Rouwenhorst (1994) predict consumption growth using data from a number of countries. Hamilton and Kim (2002) reexamine the predictability of economic activity using decomposed term yield spread for various prediction horizons (one to 16 quarters). Lint and Stolin (2003) reaffirm that economic activity is predicted by the lagged term yield spread and provide theoretical explanation for that. Another well-studied relationship in economics is the Phillips Curve, which posits that inflation is a function of unemployment. Stock and Watson (1999) predict the inflation rate by lagged unemployment rate, which is autoregressive.

Other studies deal with related predictive regression models. Jansson and Moreira (2006) provide a methodology for conducting inference on the predictive regression coefficient (with a predictor that is potentially autoregressive of order greater than 1), without providing a corresponding estimate of the predictive coefficient. Similarly, without providing point estimates, Campbell and Dufour (1997) develop a nonparametric test of the null hypothesis of no predictability in a very general context. In the context of an AR(1) predictor and a single predictive lag, Eliazs (2005) develops a a median unbiased estimator of the slope coefficient which he finds to perform well in the class of near nonstationary predictors. Chen and Deo (2009), restricting to the case of one predictive lag but allowing for a multivariate predictor, employ the Restricted Maximum Likelihood estimation and the corresponding Bartlett corrected likelihood ratio test, which they show to produce efficient and well-sized results, with higher power than that in the Jansson-Moreira (2006) test, and smaller bias than in ARM(1). Cavanagh, Elliott and Stock (1995) develop inferential methods for predictive models with a nearly-integrated predictor and a single lagged regressor in the predictive equation, whereas we deal with multi-lag predictors.

This paper proceeds as follows. Section II develops the theoretical model estimation of the slope coefficient and hypothesis testing for a model where the predictor variable has AR(p) structure with $p \geq 1$. Section III presents empirical implementation. First, we examine the usefulness of our proposed methodology in simulations. Second, we apply our method to estimate a predictive model of quarterly stock returns, the predictor being lagged dividend yield which apparently has an AR(2) structure. Using our method results in the AR(2) predictor series having insignificant effect, although under OLS the predictive model is significant. In Section IV we generalize our method to the case of multiple $AR(p)$ predictors. Our concluding remarks are in Section V.

II Augmented Regression Method (ARM) for AR(p) Predictor

We consider the following AR(p)-predictor model, for $p \geq 1$.

$$y_t = \alpha + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + u_t \quad , \tag{1}$$

$$x_t = \theta + \rho_1 x_{t-1} + \rho_2 x_{t-2} \cdots + \rho_p x_{t-p} + v_t \quad , \tag{2}$$

where $(u_t, v_t)'$ is serially independent and bivariate normal, i.e.

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \overset{i.i.d.}{\sim} N(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} \qquad .$$

This model assumes potential predictive power of *all p lags* of x_t, instead of only x_{t-1}. The problem with this predictive model is that the OLS-estimated slope coefficients $\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_p$ are potentially biased in finite samples. Stambaugh (1999) analyzes the bias in the case of $p = 1$. Our analysis proposes a method for reduced-bias parameter estimation and hypothesis testing for the general case of $p \geq 1$.

Estimating the parameters in a model such as (1) is necessary in structural models in economics and finance where, beyond the question of predictability, researchers are interested in the magnitudes of the predictive coefficients. For example, Bernanke and Mihov (1998a, 1998b) present a model, based on Bernanke and Blinder (1992), where macroeconomic variables, such as output or aggregate prices, are functions of lagged policy variables, such as money supply, with several lags. Here, policy makers need to estimate the evolving *cumulative* effect of monetary policy on output or prices. In their model, the policy variables have $AR(p)$ structure, and the macroeconomic variables are affected by p lags of the policy variables. Bernanke and Mihov (1998, p. 875) propose to "estimate [this system] by standard methods." However, if policymakers want to know the effect of policy shocks on macroeconomic variables, estimation by standard methods may produce biased coefficients, i.e., incorrect estimates of the cumulative policy effect on the economy.

In finance, point forecasts of stock returns from predictive regressions are used in constructing an optimal portfolio. For example, Lynch and Tan's (2009) predictive model uses the lagged dividend yield and the lagged book-to-market ratio (both being persistent) to predict expected returns. The point forecasts from these models are inputs in the solution of the multi-period dynamic individual portfolio choice problem. It follows that beyond knowing whether expected return is predictable by lagged regressors, this study requires point estimates of the predictive coefficients to produce the point forecast of the dependent variable.

We assume that the autoregressive model in (2) is stationary. This assumption is reasonable for a variety of applications of predictive regressions in Finance. For example, Santos and Veronesi (2006) argued that "the restriction that [the log income/consumption ratio] is stationary rests on solid economic intuition: it is not reasonable to assume that consumption can grow to be infinitely larger than labor income, or, alternatively, that labor income can grow to be several times higher than consumption."

In the model (1), (2) with $p > 1$ it is important to include all p lags in the regression. Failure to do so could result in a hypothesis test for predictability with extremely low asymptotic power. Suppose, for example, that $p = 2$, $\beta_1 = 0$, $\beta_2 \neq 0$ and $\rho_1 = 0$, so that the process has a lag-1 autocorrelation of 0. Then the OLS regression coefficient of y_t on x_{t-1} alone will converge in probability to zero, in spite of the presence of predictability of y_t. More generally, suppose in the model (1), (2) that

$$\beta_1 + \frac{Cov(x_{t-1}, (\beta_2 x_{t-2} + ... + \beta_p x_{t-p}))}{Var(x_t)} = 0 \tag{3}$$

and not all β_i with $i > 1$ are zero. Then once again there will be return predictability but regression in the misspecified predictive model for y_t based on x_{t-1} alone will produce an estimated predictive coefficient of zero, asymptotically. Finally, even if the lefthand side of (3) is nonzero but is sufficiently close to zero, the misspecified regression of y_t on x_{t-1} alone could lead to extremely low finite-sample power in the corresponding test for return predictability.[3]

A Reduced-bias estimation of predictive coefficients

We outline our proposed procedure to produce reduced-bias slope coefficients. In model (1), u_t can be decomposed into

$$u_t = \phi v_t + e_t \tag{4}$$

where $\{e_t\}$ are $i.i.d.$ normal and independent of both $\{v_t\}$ and $\{x_t\}$. It is easy to see that $\phi = \sigma_{uv}/\sigma_v^2$. We then construct a proxy $\{v_t^c\}$ for $\{v_t\}$,

$$v_t^c = x_t - \hat{\theta}^c - \hat{\rho}_1^c x_{t-1} - \hat{\rho}_2^c x_{t-2} \cdots - \hat{\rho}_p^c x_{t-p} \quad , \tag{5}$$

where $\hat{\theta}^c$, $\hat{\rho}_1^c, \cdots, \hat{\rho}_p^c$ are estimators of θ, $\rho_1, \cdots, \rho_p$ based on the available data, $\{x_t\}_{t=-p+1}^n$. Specific choices for these estimators are given below.

Our reduced-bias predictive coefficients $\hat{\beta}_i^c, i = 1, 2, \ldots, p$ are produced by an *augmented regression*, where $\{y_t\}_{t=1}^n$ is regressed by OLS on $\{x_{t-1}\}_{t=1}^n, \{x_{t-2}\}_{t=1}^n, \ldots, \{x_{t-p}\}_{t=1}^n$ and on $\{v_t^c\}_{t-1}^n$, with intercept.

[3] In both our model and the augmented regression estimation method presented here, we assume that the number of lags in (1) and (2) are the same. This raises a model selection problem which is beyond the scope of this paper. Above, we discussed the consequences of omitting a relevant predictive variable in (1). If, however, we estimate an irrelevant variable in (1), for example, if $\beta_2 = 0$ but we include x_{t-2} in the regression, then asymptotically there is no harm done, but in finite samples it will entail some cost in terms of both efficiency and power.

Theorem 1 *The bias of* $\hat{\beta}_i^c, (i = 1, \ldots, p)$ *is given by*

$$E[\hat{\beta}_i^c - \beta_i] = \phi E[\hat{\rho}_i^c - \rho_i] \quad .$$

Proof: See appendix.

Following Theorem 1, the bias of the predictive coefficients is reduced if the estimators $\hat{\rho}_j^c, (j = 1, \ldots, p)$ are selected to be as nearly unbiased as possible for ρ_j. This follows since the bias of $\hat{\beta}_j^c$ is proportional to the bias of $\hat{\rho}_j^c$, as in Theorem 1. This result is a generalization of the bias expression in Stambaugh (1999) for the AR(1) predictor case.

The estimated coefficient $\hat{\phi}^c$ of v_t^c obtained from the augmented regression is unbiased:

Lemma 1 $E[\hat{\phi}^c] = \phi$.

Proof: See appendix.

Combining Equations (1) (2), (4) and (5) we have

$$\begin{aligned} y_t &= \alpha + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + \phi v_t + e_t \\ &= \alpha + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + \phi(v_t - v_t^c) + \phi v_t^c + e_t \\ &= [\alpha + \phi(\hat{\theta}^c - \theta)] + [\beta_1 + \phi(\hat{\rho}_1^c - \rho_1)]x_{t-1} + \cdots + [\beta_p + \phi(\hat{\rho}_p^c - \rho_p)]x_{t-p} + \phi v_t^c + e_t \end{aligned}$$

where the error terms e_t are i.i.d. normal with mean zero, and for all t, e_t is independent of $x_{-p+1}, \ldots, x_n$.

The unbiasedness of $\hat{\phi}^c$ is seen from the fact that, conditionally on $x_{-p+1}, \ldots, x_n$, the model above satisfies the usual regularity conditions for a regression model (such as independence between the error term and all regressors, including v_t^c, which is a function of $x_{-p+1}, \ldots, x_n$), based on the full set of regressors, that is, all p lags of x_t together with v_t^c.

Bias-corrected estimators of ρ_i can be obtained from Shaman and Stine's (1988, 1989) small-sample bias expressions for the OLS estimates $\hat{\rho}_i$ in an AR(p) process. For example, for the $p = 2$ case, (See the Appendix for the cases $p = 1, \ldots, 5$)

$$\begin{aligned} E(\hat{\rho}_1 - \rho_1) &= -\frac{1 + \rho_1 + \rho_2}{n} \\ E(\hat{\rho}_2 - \rho_2) &= -\frac{2 + 4\rho_2}{n} \ . \end{aligned}$$

We use these expressions to construct bias-corrected estimators $\hat{\rho}_1^c, \cdots, \hat{\rho}_p^c$ of the OLS estimates $\hat{\rho}_1, \cdots, \hat{\rho}_p$. For any AR(p) model, the bias expressions are linear functions of the true autocorrelations $\rho_1, \cdots, \rho_p$. Plugging the OLS estimators $\hat{\rho}_1, \cdots, \hat{\rho}_p$ into these expressions, and then subtracting the result from the corresponding $\hat{\rho}_i$ yields the reduced-bias estimators we will use, $\hat{\rho}_1^c, \cdots, \hat{\rho}_p^c$. For example, with $p = 2$, we have

$$\begin{aligned} \hat{\rho}_1^c &= \hat{\rho}_1 + \frac{1 + \hat{\rho}_1 + \hat{\rho}_2}{n} \\ \hat{\rho}_2^c &= \hat{\rho}_2 + \frac{2 + 4\hat{\rho}_2}{n} \ . \end{aligned}$$

B Hypothesis testing: Estimating $\mathrm{cov}(\hat{\beta}_i^c, \hat{\beta}_j^c)$

Having estimated the reduced-biased coefficients of the predictor variable, $\hat{\beta}_i^c$, we propose a method to test hypotheses related to these coefficients. We use the following feasible formulas to estimate the covariance between $\hat{\beta}_i^c$ and $\hat{\beta}_j^c$, motivated by Lemma 2 and (8) below,

$$\widehat{\mathrm{cov}}^c(\hat{\beta}_i^c, \hat{\beta}_j^c) = \{\hat{\phi}^c\}^2 \widehat{\mathrm{cov}}(\hat{\rho}_i^c, \hat{\rho}_j^c) + \widehat{\mathrm{cov}}(\hat{\beta}_i^c, \hat{\beta}_j^c) \tag{6}$$

and for the case $(i = j)$,

$$\widehat{\mathrm{var}}^c(\hat{\beta}_i^c) = \{\hat{\phi}^c\}^2 \widehat{\mathrm{var}}(\hat{\rho}_i^c) + \widehat{\mathrm{var}}(\hat{\beta}_i^c) \ . \tag{7}$$

To evaluate the covariance, we need the following theoretical results.

Since $E[(\hat{\beta}_i^c-\beta_i)(\hat{\beta}_j^c-\beta_j)] = \text{cov}[\hat{\beta}_i^c, \hat{\beta}_j^c] + E[\hat{\beta}_i^c-\beta_i]\cdot E[\hat{\beta}_j^c-\beta_j]$, and since by Theorem 1, the bias is $E[\hat{\beta}_i^c-\beta_i] = \phi E[\hat{\rho}_i^c-\rho_i] = O(1/n^2)$, we obtain

$$\text{cov}[\hat{\beta}_i^c, \hat{\beta}_j^c] = E[(\hat{\beta}_i^c-\beta_i)(\hat{\beta}_j^c-\beta_j)] + O(1/n^4) \tag{8}$$

where the first term can be evaluated using Lemma 2.

Lemma 2

$$E[(\hat{\beta}_i^c-\beta_i)(\hat{\beta}_j^c-\beta_j)] = \phi^2 E[(\hat{\rho}_i^c-\rho_i)(\hat{\rho}_j^c-\rho_j)] + E[\widehat{\text{cov}}(\hat{\beta}_i^c, \hat{\beta}_j^c)] \tag{9}$$

where $\widehat{\text{cov}}(\hat{\beta}_i^c, \hat{\beta}_j^c)$ is the estimated covariance between $\hat{\beta}_i^c$ and $\hat{\beta}_j^c$, based on an OLS regression of y_t on $x_{t-1}, \cdots, x_{t-p}$ and v_t^c, with intercept (provided by standard regression packages).

Proof: See appendix.

If $i=j$, the above lemma simplifies to

$$E[\hat{\beta}_i^c-\beta_i]^2 = \phi^2 E[\hat{\rho}_i^c-\rho_i]^2 + E[\widehat{\text{var}}(\hat{\beta}_i^c)]. \tag{10}$$

We now need to accurately estimate $\phi^2 E[(\hat{\rho}_i^c-\rho_i)(\hat{\rho}_j^c-\rho_j)]$. First, we note that the coefficient $\hat{\phi}^c$ of v_t^c in the augmented regression is unbiased (see Lemma 1 above). Next, we need to construct an estimator of $E[(\hat{\rho}_i^c-\rho_i)(\hat{\rho}_j^c-\rho_j)]$ with low bias. Here we use some heuristic approximations (as in Amihud and Hurvich (2004)), which turn out to work quite well in simulations. It follows from Shaman and Stine (1988) that $\hat{\rho}_i^c$ is a low-bias estimator of ρ_i with bias that is $O(1/n^2)$. We therefore treat the autoregressive coefficients $\hat{\rho}_i^c$ as if they were unbiased. Then we simply need an expression for $\text{cov}(\hat{\rho}_i^c, \hat{\rho}_j^c)$, which can be easily obtained because all plug-in versions of the bias correction for $\hat{\rho}_i^c$ can be expressed as a linear function of $\hat{\rho}_j, (j = 1, \ldots, p)$, according to Shaman and Stine (1988).

In the particular case $p = 2$, the Shaman and Stine (1988) corrections are given in section II.A. Based on them, feasible approximations for $\text{var}(\hat{\rho}_1^c)$, $\text{var}(\hat{\rho}_2^c)$ and $\text{cov}(\hat{\rho}_1^c, \hat{\rho}_2^c)$ are given by

$$\begin{aligned}
\widehat{\text{var}}(\hat{\rho}_1^c) &= (1 + \frac{1}{n})^2 \widehat{\text{var}}(\hat{\rho}_1) + \frac{1}{n^2} \widehat{\text{var}}(\hat{\rho}_2) + 2(1 + \frac{1}{n})(\frac{1}{n}) \widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2) \\
\widehat{\text{var}}(\hat{\rho}_2^c) &= (1 + \frac{4}{n})^2 \widehat{\text{var}}(\hat{\rho}_2) \\
\widehat{\text{cov}}(\hat{\rho}_1^c, \hat{\rho}_2^c) &= (\frac{1}{n})(1 + \frac{4}{n}) \widehat{\text{var}}(\hat{\rho}_2) + (1 + \frac{1}{n})(1 + \frac{4}{n}) \widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2)
\end{aligned}$$

where $\widehat{\text{var}}(\hat{\rho}_1)$, $\widehat{\text{var}}(\hat{\rho}_2)$ and $\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2)$ are obtained from the OLS regression of model (2). Using these, together with (6) and (7), we can estimate the covariance $\text{cov}(\hat{\beta}_1^c, \hat{\beta}_2^c)$.

Lemma 3

$$\text{var}(\hat{\phi}^c) = E[\widehat{\text{var}}(\hat{\phi}^c)] \tag{11}$$

where $\widehat{\text{var}}(\hat{\phi}^c)$ is the estimated standard error for $\hat{\phi}^c$ as provided by standard regression packages, based on an OLS regression of y_t on $x_{t-1}, \ldots, x_{t-p}$ and v_t^c with intercept.

III Implementation

The ARM(p) estimation procedure can be summarized as follows:

(i) Estimate model (2) by OLS,

(ii) Apply Shaman and Stine's (1988, 1989) reduced-bias estimators of ρ_i to obtain

v_t^c, given by (5). For example, for the $p = 2$ case, the plug-in versions for $\hat{\rho}_i^c$ are

$$\begin{aligned} \hat{\rho}_1^c &= \hat{\rho}_1 + \frac{1 + \hat{\rho}_1 + \hat{\rho}_2}{n} \\ \hat{\rho}_2^c &= \hat{\rho}_2 + \frac{2 + 4\hat{\rho}_2}{n} \end{aligned}$$

(iii) Perform an augmented OLS regression of y_t on $x_{t-1}, \ldots, x_{t-p}$ and v_t^c (with intercept) to obtain reduced-bias estimates $(\hat{\beta}_1^c, \ldots, \hat{\beta}_p^c)$.

The hypothesis testing procedure is summarized as follows:

(i) Estimate the $(p \times p)$ covariance matrix for the vector $(\hat{\beta}_1^c, \ldots, \hat{\beta}_p^c)'$ by Equation (6) and (7). Denote this estimated covariance matrix by $\hat{\Gamma}_\beta$.

(ii) Using $\hat{\Gamma}_\beta$ together with $(\hat{\beta}_1^c, \ldots, \hat{\beta}_p^c)'$, perform individual t-tests for $\beta_i, (i = 1, \ldots, p)$ based on the statistic $\hat{\beta}_i^c / \sqrt{\hat{\Gamma}_\beta(i,i)}$ where $\hat{\Gamma}_\beta(i,i)$ is the i-th diagonal element of $\hat{\Gamma}_\beta$.

(iii) A Wald-type joint test can be constructed similarly using $\hat{\Gamma}_\beta$ and $(\hat{\beta}_1^c, \ldots, \hat{\beta}_p^c)'$.

A Simulation Study

We investigate the performance of parameter estimation and hypothesis testing using ARM(p) in a simulation study, using 10000 simulated replications from the model (1) and (2). Specifically, we compare OLS and ARM(p) method in terms of the bias in estimating the predictive coefficients $\beta = (\beta_1, \ldots, \beta_p)$ and in terms of the size of the statistical tests on hypothesis tests for the β coefficients. We expect that the under ARM(p), the bias is smaller and the tests are more accurate than under OLS. This is in fact what we obtain.

The simulation results are presented in Table 1. We assume an AR(2) predictor model and do the simulations for two sample sizes, $n = 50$ and $n = 200$. Naturally, we expect that the bias under OLS is greater for the smaller n and therefore it is for the smaller sample size that our ARM(p) provides greater improvement. The parameter values that we use are obtained from an empirical analysis, presented in the next section, of predicting the quarterly NYSE stock returns by the dividend/price ratio. Accordingly, we set $\rho_1 = 1.1053$, $\rho_2 = -0.1430$, $\phi = -92.17$, $\sigma_e = 0.01844$ and $\sigma_v = 0.0007746$ as Case 1. The corresponding roots for the AR(2) process are 0.9557 and 0.1496. To examine our method for an AR(2) process whose highest root is lower, we set $\rho_1 = 1.0553$ while holding $\rho_2 = -0.1430$, in which case the corresponding roots are 0.8956 and 0.1597. This is Case 2. All other parameters are the same, including the setting of $\beta_1 = \beta_2 = 0$. The table shows the parameter estimates under both OLS and ARM(p) and the realized size under a 5% nominal size used for both the t-tests—right-sided and two-sided—and the Wald test.

INSERT TABLE 1 HERE

We find that the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are more biased than the ARM(p) estimators $\hat{\beta}_1^c$ and $\hat{\beta}_2^c$. In Case 1 (with $\rho_1 = 1.1053$), for $n = 200$, averaging over 10000 realizations, we obtain $(\hat{\beta}_1, \hat{\beta}_2) = (1.030, 0.657)$, while the true values are (0,0). Under ARM(p), the bias is much smaller: $(\hat{\beta}_1^c, \hat{\beta}_2^c) = (0.134, 0.013)$. For $n = 50$, the reduction of the OLS bias is greater under ARM(p): the estimated bias declines from (5.263, 2.030) under OLS to (1.793, -0.439) under ARM(2). The variance of $\hat{\beta}_i^c, (i = 1, 2)$ is slightly larger than that of $\hat{\beta}_i, (i = 1, 2)$, hence the root mean squared errors (RMSE) of the ARM(p) slope estimates—which includes the effect of the bias—are in some cases greater than those under OLS. Nevertheless, *the ARM(p)-based hypothesis testing always produces more accurate sizes than OLS-based hypothesis testing.* The parameter estimates for Case 2—with a lower highest root—are qualitatively similar.

Test results—comparisons the realized sizes with the nominal size of 5%—are reported in Table 1, Panel B, for right-sided and two-sided hypothesis tests.[4] The null hypothesis is $H_0 : \beta_1 = 0$ for $t(\hat{\beta}_1)$; $H_0 : \beta_2 = 0$ for $t(\hat{\beta}_2)$ and $H_0 : \beta_1 = \beta_2 = 0$ for the joint Wald test. For the individual coefficient tests we employ standard t-values (corresponding to 5% tail probability), and for the joint test we employ a Wald test with standard values of χ^2 that would reject the null hypothesis if it were true 5% of the time. We then report the frequency at which the null is actually rejected when it is true—this is the realized size. Consider Case 1 with $n = 50$. For the two-tailed test, the realized sizes for the tests of $(\hat{\beta}_1)$ and $(\hat{\beta}_2)$ are 9.5% and 5.7%, respectively, while ARM(p)-based tests produce smaller realized sizes, 7.7% and 5.1%, for $(\hat{\beta}_1^c)$ and $(\hat{\beta}_2^c)$, respectively. For the right-tail test, the improvements in the realized sizes is greater, declining from 12.5% and 7.3% under OLS for the two slope coefficients to 8.2% and 4.9% under ARM(p). The ARM-based test thus greatly improves the size (makes it closer to the nominal size). That is, the null hypothesis is rejected under OLS tests more often than it should be, and more often than it is rejected under ARM-based tests. Still, the ARM-based test sometimes results in too large a size for $(\hat{\beta}_1^c)$, reflecting the impact of high value of ρ_1 and consequently the high value of the largest root, which is close to unity, in which case the process is close to being non-stationary. An alternative test could be performed by generating simulation-based critical values for the hypothesis testing instead of using those based on percentiles of the normal distribution.[5] For the larger sample size, $n = 200$, the realized sizes are naturally closer to the nominal sizes and ARM(p) produces sizes which are quite accurate, outperforming those under OLS. Similar patterns are observed for the Wald test. Finally, for Case 2 where the highest root is lower, the realized size are closer to the nominal size compared to those in Case 1 where the process is closer to non-stationarity, and again they are less distorted than those under OLS. The improvement of ARM(p) over OLS is

[4] The results for nominal sizes 1% and 10% are qualitatively similar.

[5] For detailed discussion of the computation of these critical values, see Amihud, Hurvich and Wang (2006).

again greater for the smaller sample size (n=50).

B Empirical Analysis

We study the prediction of $VWNY$, the NYSE value-weighted stock return, by the popular predictor dividend yield, DY. Return is quarterly and DY pertains to the end of the quarter, where dividend is summed over the past year and divided by the end-of-quarter price.[6] The study period is 1946-1994 during which DY has been shown to have stronger predictive power than it has when adding the period 1995-2000, "during which time prices moved strongly against the predictions of the model" (Lewellen (2004, p.224)).[7] During 1946-1994, DY significantly predicts *monthly* stock returns even after accounting for the AR(1)-induced bias in the predictive coefficient. We show that DY significantly predicts *quarterly* stock returns using OLS, which is known to produce biased predictive coefficient (Stambaugh (1999)). When we employ the standard correction for an assumed AR(1) structure of quarterly DY, its predictive power is still significant. However, our estimation shows that the correct autoregressive structure of quarterly DY is AR(2) rather than AR(1). Then, using our augmented regression method for AR(2), the predictive power of DY becomes insignificant.

The predictor series $\log DY$ is identified to be an AR(2) by Akaike's (1974) information criterion. We therefore employ the following estimation model (Model A):

$$\begin{aligned} VWNY_t &= \alpha + \beta_1 \log(DY_{t-1}) + \beta_2 \log(DY_{t-2}) + u_t \\ \log(DY_t) &= \theta + \rho_1 \log(DY_{t-1}) + \rho_2 \log(DY_{t-2}) + v_t \end{aligned}$$

INSERT TABLE 2 HERE

[6] Data are kindly provided by Jon Lewellen.

[7] See also Ang and Bekaert (2007, Table 2).

The estimation results, presented in Table 2, show that the appropriate autoregressive model for $\log DY_t$ is AR(2), with both autoregressive coefficients being statistically significant. The largest root, given our estimate of $\hat{\rho}_1^c = 1.1053$ and $\hat{\rho}_2^c = -0.1430$ is 0.9557. For testing the joint predictive effect of $\log DY_{t-1}$ and $\log DY_{t-2}$ we employ the Wald test, since the high correlation between the two predictors makes individual t-tests inappropriate here. The Wald test is of the *joint* effect of $\log DY_{t-1}$ and $\log DY_{t-2}$, that is, whether the vector (β_1, β_2) is significantly different from (0,0). The OLS Wald test results shows that $\log DY_{t-1}$ and $\log DY_{t-2}$ jointly predict $VWNY$ with high statistical significance. However, when employing the ARM(2) test, the joint Wald test shows that there is no significant predictive effect of lagged dividend yield. The value of the Wald test statistic is 3.91 while the critical value for 5% significance is 5.99. Note that, even though the test is over-sized (i.e., it rejects the null too often, as we show in the simulations), we still fail to reject the null when using the ARM-based test. The p-value under the ARM-based Wald test, 0.142, is much greater (177 times greater) than the p-value of 0.0008 under the OLS-based Wald test. That is, the OLS would lead to a sound rejection of the null, implying significant predictability while there is none.[8]

Most existing predictive regression literature considers the case the predictor series is first-order autoregressive, AR(1), which is appropriate for some data. But researchers do not always investigate the exact autoregressive structure of the predictor variable series. We estimate a predictive regression model assuming that $\log DY$ is AR(1) (as it apparently is in monthly data). This implies the following predictive regression model (Model B):

$$
\begin{aligned}
VWNY_t &= \alpha + \beta_1 \log DV_{t-1} + u_t \\
\log DV_t &= \theta + \rho_1 \log DV_{t-1} + v_t
\end{aligned}
$$

[8] It is worth noting, however, that the power of the ARM(p)-based test may be reduced due to altering the specification of the predictive regression model.

When this model is estimated by OLS, we obtain $\hat{\beta}_1 = 0.0728$ with $t = 3.43$, highly significant. The AR(1) coefficient of $\log DY_{t-1}$ is $\hat{\rho}_1^c = 0.9729$.[9] To correct for the well-known bias in this case, we apply the Amihud-Hurvich (2004) method assuming AR(1). This produces a reduced-bias estimate of $\hat{\beta}_1^c = 0.0546$ with $t = 2.53$, still statistically significant. However, as we have argued above, the autoregressive structure of $\log DY$ is apparently AR(2) and not AR(1). When employing ARM(2), the predictive power of $\log DY$ becomes insignificant.

IV Extension: Multiple $AR(p)$ Predictors

Here, we generalize the $ARM(p)$ method to the case of multiple predictors in which each predictor is $AR(p)$, with $p \geq 1$. For notational simplicity, we assume that each predictor has the same autoregressive order. Suppose, then, that $\{x_t\}$ is a stationary q-dimensional series, and that we wish to use p lags of x_t to predict the univariate response y_t. We assume that $\{x_t\}$, $\{y_t\}$ are given by the model

$$y_t = \alpha + \beta_1' x_{t-1} + \cdots + \beta_p' x_{t-p} + u_t \tag{12}$$

$$x_t = \Theta + \Phi_1 x_{t-1} + \cdots + \Phi_p x_{t-p} + v_t \tag{13}$$

$$u_t = \phi' v_t + e_t. \tag{14}$$

In (12), we assume that $\{y_t\}$, α, and $\{u_t\}$ are (1×1), β_i are $(q \times 1)$ for $i = 1, \cdots, p$, and $\{x_t\}$ is $(q \times 1)$. In (13), we assume that $\{v_t\}$ is a $(q \times 1)$ Gaussian white noise series with $cov(v_t) = \Sigma_v$, Θ is $(q \times 1)$, Φ_i are $(q \times q)$ and diagonal for $i = 1, \cdots, p$, and that $\{x_t\}$ is stationary. In (14), we assume that ϕ is $(q \times 1)$, the $\{e_t\}$ are independent and identically distributed normal with mean zero, and that $\{e_t\}$ is independent of both $\{v_t\}$ and $\{x_t\}$.

[9] The estimate $\hat{\rho}^c$ suggests that using Lewellen's (2004) method of setting $\rho = 0.9999$ is inappropriate here.

Combining (12) and (14), we obtain

$$y_t = \alpha + \beta_1' x_{t-1} + \cdots + \beta_p' x_{t-p} + \phi' v_t + e_t. \tag{15}$$

We can construct a proxy $\{v_t^c\}$ for $\{v_t\}$, given by

$$v_t^c = x_t - \hat{\Theta}^c - \hat{\Phi}_1^c x_{t-1} - \cdots - \hat{\Phi}_p^c x_{t-p} \tag{16}$$

where $\hat{\Theta}^c$, $\hat{\Phi}_i^c$, are estimators of θ, Φ_i, based on the available data, $\{x_t\}_{t=-p+1}^n$. Specific choices for these estimators are given below. We will assume $\hat{\Phi}_i^c$ to be diagonal.

Our reduced-bias predictive coefficient vectors $\hat{\beta}_i^c, i = 1, 2, \ldots, p$, are obtained by an augmented OLS regression of $\{y_t\}_{t=1}^n$ on all pq entries of $\{x_{t-1}\}_{t=1}^n, \{x_{t-2}\}_{t=1}^n, \ldots, \{x_{t-p}\}_{t=1}^n$ as well as all entries of $\{v_t^c\}_{t=1}^n$, with intercept. The j^{th} entry of $\hat{\beta}_i^c$, denoted by $\hat{\beta}_{i,j}^c$ is the coefficient of the j^{th} entry of $\{x_{t-i}\}$ in this regression for $i = 1, \cdots, p$, $j = 1, \cdots, q$. Here, the i, j subscript refers to the i^{th} lag, j^{th} variable. We also obtain the estimators $\hat{\phi}_j^c$ as the coefficient of the j^{th} entry of v_t^c in the regression, for $j = 1, \cdots, q$. The $\hat{\phi}_j^c$ are estimators of the j^{th} entry of ϕ.

Following along the lines of the proofs of Theorem 1 and Lemma 1, we obtain the following results (omitting the proofs for the sake of brevity).

Theorem 2 *The bias of $\hat{\beta}_i^c, (i = 1, \ldots, p)$ is given by*

$$E[\hat{\beta}_i^c - \beta_i] = E[\hat{\Phi}_i^c - \Phi_i]'\phi \quad .$$

Lemma 4 $E[\hat{\phi}^c] = \phi$.

As seen from Theorem 2, since Φ_i and $\hat{\Phi}_i^c$ are diagonal, the bias in the j^{th} entry of $\hat{\beta}_i^c$ is proportional to the bias in $\hat{\rho}_{i,j}^c$ as an estimator of $\rho_{i,j}$, where $\hat{\rho}_{i,j}^c$ is the j^{th} diagonal entry of

$\hat{\Phi}_i^c$ and $\rho_{i,j}$ is the j^{th} diagonal entry of Φ_i. Using this notation, the j^{th} predictor variable $x_{t,j}$ is autoregressive of order p, given by $x_{t,j} = \Theta_j + \rho_{1,j}x_{t-1,j} + \cdots + \rho_{p,j}x_{t-p,j} + v_{t,j}$. For each j, we can therefore focus on the j^{th} predictor variable alone, and obtain the bias corrected estimators $\hat{\rho}_{1,j}^c, \cdots, \hat{\rho}_{p,j}^c$ by the univariate $AR(p)$ method described earlier, using the bias expressions of Shaman and Stine (1988, 1989). We then obtain

$$\hat{\Theta}_j^c = \frac{1}{n}\sum_{t=1}^{n}(x_{t,j} - \hat{\rho}_{1,j}^c x_{t-1,j} - \cdots - \hat{\rho}_{p,j}^c x_{t-p,j}).$$

A similar argument to the one leading to (6) provides motivation for the following estimator of the covariance between the entries of the reduced bias predictive coefficient vectors, *viz.*,

$$\widehat{\text{cov}}^c(\hat{\beta}_{i_1,j_1}^c, \hat{\beta}_{i_2,j_2}^c) = \hat{\phi}_{j_1}^c \hat{\phi}_{j_2}^c \widehat{cov}(\hat{\rho}_{i_1,j_1}^c, \hat{\rho}_{i_2,j_2}^c) + \widehat{cov}(\hat{\beta}_{i_1,j_1}^c, \hat{\beta}_{i_2,j_2}^c) \tag{17}$$

for $i_1, i_2 = 1, \cdots, p$ and $j_1, j_2 = 1, \cdots, q$, where $\widehat{cov}(\hat{\rho}_{i_1,j_1}^c, \hat{\rho}_{i_2,j_2}^c)$ is defined below and $\widehat{\text{cov}}(\hat{\beta}_{i_1,j_1}^c, \hat{\beta}_{i_2,j_2}^c)$ is the estimated covariance between $\hat{\beta}_{i_1,j_1}^c$ and $\hat{\beta}_{i_2,j_2}^c$, based on the augmented OLS regression (provided by standard regression packages).

We now explain how to obtain $\widehat{cov}(\hat{\rho}_{i_1,j_1}^c, \hat{\rho}_{i_2,j_2}^c)$. Since the bias-corrected estimators of the autoregressive parameters are linear combinations of the OLS estimators, it suffices to construct an estimator of the covariance matrix $\text{cov}(\hat{\rho})$, where $\hat{\rho}$ is the OLS estimator of $\rho = (\rho_{1,1}, \cdots, \rho_{p,1}, \cdots, \rho_{1,q}, \cdots, \rho_{p,q})'$. Since the Φ_i matrices are diagonal, the OLS estimators $\hat{\rho}_{1,j}, \cdots, \hat{\rho}_{p,j}$ can be obtained directly from the $AR(p)$ equation for the j^{th} variable, but since we will also need covariances between OLS estimators corresponding to different variables, it is helpful to express $\hat{\rho}$ as a subvector of $(X'X)^{-1}X'\tilde{y}$ where $\tilde{y} = (x_{1,1}, \cdots, x_{n,1}, \cdots, x_{1,q}, \cdots, x_{n,q})'$ and X is an appropriately chosen $(nq \times (nq + q))$

matrix such that $X'X$ is block diagonal. For example, in the case $p = q = 2$, we have

$$X = \begin{pmatrix} 1 & x_{0,1} & x_{-1,1} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n-1,1} & x_{n-2,1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{0,2} & x_{-1,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & x_{n-1,2} & x_{n-2,2} \end{pmatrix}$$

If $\tilde{v} = (v_{1,1}, \cdots, v_{n,1}, \cdots, v_{1,q}, \cdots, v_{n,q})'$ then $cov(\hat{\rho})$ is a submatrix of

$$(X'X)^{-1}X'Cov(\tilde{v})X(X'X)^{-1}.$$

The model assumptions imply that the entries of $Cov(\tilde{v})$ are determined by Σ_v, which we estimate by $\hat{\Sigma}_v = \frac{1}{n-1}\sum_{t=1}^{n} v_t^c(v_t^c)'$. This determines $\widehat{cov}(\hat{\rho})$. Then the vector of bias corrected estimators $\hat{\rho}^c$ of ρ is constructed as a product of a fixed matrix with $\hat{\rho}$, leading to $\widehat{cov}(\hat{\rho}^c)$ and ultimately the desired values $\widehat{cov}(\hat{\rho}^c_{i_1,j_1}, \hat{\rho}^c_{i_2,j_2})$ for use in (17). The estimated covariance matrix determined by $\widehat{\mathrm{cov}}^c(\hat{\beta}^c_{i_1,j_1}, \hat{\beta}^c_{i_2,j_2})$ from (17) can be used to construct a Wald test of the joint null hypothesis that all predictive coefficients are zero.

We ran simulations for the case $p = q = 2$, all with $\phi_1 = \phi_2 = -92.17$, $\beta_1 = \beta_2 = (0, 0)'$, $\alpha = 0$, $\Theta = (0, 0)'$, $var(e_t) = 0.01844^2$ and

$$\Sigma_v = (0.0007746)^2 \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$

We considered two different models. In the first model, we took $\Phi_1 = diag(1.1053, 1.1053)$, $\Phi_2 = diag(-0.144, -0.144)$. In the second model, we took $\Phi_1 = diag(1.0553, 1.0553)$, $\Phi_2 = diag(-0.144, -0.144)$. For each model, we ran 2,000 replications, for the two sample sizes $n = 50$ and $n = 200$. We report here only some of the results we obtained for the first model. See Table 3.

INSERT TABLE 3 HERE

It is seen from Table 3 that the ARM-based estimates of the β parameters have not only dramatically lower bias, but also much smaller standard deviations (by a factor of approximately 2) compared with the corresponding OLS estimates. The smaller standard deviation implies that the power of ARM-based tests for predictability would be higher than for the corresponding OLS-based tests. The ARM-based t-tests and Wald test are less oversized than the corresponding OLS-based tests, though the ARM-based t-tests are still noticeably oversized for $n = 50$, and the ARM-based Wald test is noticeably oversized for both sample sizes, in the situation studied.

V Conclusion

This paper emphasizes that in predictive regressions, where one variable is predicted by lagged values of another variable, it is important to correctly identify the autoregressive structure of the predictor variable series. Current research on predictive regressions studies the case where the predictive series is first-order autoregressive, AR(1), which is indeed appropriate for some data. We develop an augmented regression method for the case where the predictor variable is autoregressive of order p, $p \geq 1$, denoted ARM(p). It reduces to Amihud and Hurvich's (2004) method when p=1. For predictive regression with predictor series that are AR(p), we proposed bias-reduced point estimation of the predictive coefficients and a corresponding hypothesis testing procedure. We show, both theoretically and by simulations, that the use of OLS in such a model may produce biased estimates of the predictive regression coefficients. This is conceptually consistent with the analysis of Stambaugh (1999) and others for predictor series that are AR(1). Applying ARM(p) to a model where quarterly stock returns are predicted by dividend yield, we find that the predictor series is AR(2). For these data, we find that dividend yield is a significant predictor of stock returns not only based on OLS but also based on the standard bias-correction method that assumes that the predictor series is AR(1). However, the predictor series is found to be AR(2), and our ARM(2) method results in the estimated predictor coefficients being insignificantly different from zero.

VI Appendix

Proof of Theorem 1: Combining Equations (1) (2), (4) and (5) we have

$$\begin{aligned} y_t &= \alpha + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + \phi v_t + e_t \\ &= \alpha + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + \phi(v_t - v_t^c) + \phi v_t^c + e_t \\ &= [\alpha + \phi(\hat{\theta}^c - \theta)] + [\beta_1 + \phi(\hat{\rho}_1^c - \rho_1)]x_{t-1} + \cdots + [\beta_p + \phi(\hat{\rho}_p^c - \rho_p)]x_{t-p} + \phi v_t^c + e_t \end{aligned} \tag{18}$$

where the error terms e_t are i.i.d. normal with mean zero, and for all t, e_t is independent of $x_{-p+1}, \ldots, x_n$.

Let $\{r_{1,t}\}_{t=1}^n$ denote the residuals from an OLS regression of $\{x_{t-1}\}_{t=1}^n$ on $\{x_{t-2}\}_{t=1}^n$, $\cdots$, $\{x_{t-p}\}_{t=1}^n$ and $\{v_t^c\}_{t=1}^n$. It follows that

$$\sum_{t=1}^n r_{1,t} = 0 \quad , \quad \sum_{t=1}^n r_{1,t} x_{t-1} = \sum_{t=1}^n r_{1,t}^2 \quad ,$$

$$\sum_{t=1}^n r_{1,t} x_{t-2} = 0 \quad , \cdots , \quad \sum_{t=1}^n r_{1,t} x_{t-p} = 0 \quad , \quad \sum_{t=1}^n r_{1,t} v_t^c = 0 \quad . \tag{19}$$

We have

$$\hat{\beta}_1^c = \frac{\sum_{t=1}^n r_{1,t} y_t}{\sum_{t=1}^n r_{1,t}^2} \quad . \tag{20}$$

Combining (19) and (20), we obtain

$$\hat{\beta}_1^c - \beta_1 = \phi(\hat{\rho}_1^c - \rho_1) + \frac{\sum_{t=1}^n r_{1,t} e_t}{\sum_{t=1}^n r_{1,t}^2} \quad . \tag{21}$$

Since $\hat{\theta}^c$, $\hat{\rho}_1^c$, $\cdots$, $\hat{\rho}_p^c$ and $\{v_t^c\}$ are all functions of $x_{-p+1}, \cdots, x_n$, and hence are independent of $\{e_t\}$, the expectation of the last term on the righthand side of (21) is zero, so

$$E[\hat{\beta}_1^c - \beta_1] = \phi E[\hat{\rho}_1^c - \rho_1]$$

Similarly, if we let $\{r_{j,t}\}_{t=1}^n$ denote the residuals from an OLS regression of $\{x_{t-j}\}_{t=1}^n$ on $\{x_{t-1}\}_{t=1}^n$, $\{x_{t-j+1}\}_{t=1}^n$, ..., $\{x_{t-j-1}\}_{t=1}^n$, $\{x_{t-p}\}_{t=1}^n$ and $\{v_t^c\}_{t=1}^n$, we obtain

$$E[\hat{\beta}_j^c - \beta_j] = \phi E[\hat{\rho}_j^c - \rho_j] \tag{22}$$

for all $(j = 1, \ldots, p)$. □

Proof of Lemma 1: We first note that conditionally on $x_{-p+1}, \ldots, x_n$, Equation (18) satisfies all the regularity conditions needed for a linear regression model, and therefore

$$E[\hat{\phi}^c | x_{-p+1}, \ldots, x_n] = \phi \quad .$$

Taking the expectation of the formula above and applying the double expectation theorem completes the proof. □

Proof of Lemma 2: Arguing as in (21) we have

$$\begin{aligned} \hat{\beta}_i^c - \beta_i &= \phi(\hat{\rho}_i^c - \rho_i) + \frac{\sum_{t=1}^n r_{i,t} e_t}{\sum_{t=1}^n r_{i,t}^2} \\ \hat{\beta}_j^c - \beta_j &= \phi(\hat{\rho}_j^c - \rho_j) + \frac{\sum_{t=1}^n r_{j,t} e_t}{\sum_{t=1}^n r_{j,t}^2} \end{aligned}$$

so that

$$E[(\hat{\beta}_i^c - \beta_i)(\hat{\beta}_j^c - \beta_j)] = \phi^2 E[(\hat{\rho}_i^c - \rho_i)(\hat{\rho}_j^c - \rho_j)] + \sigma_e^2 E\left[\frac{\sum_{t=1}^n r_{i,t} r_{j,t}}{\sum_{t=1}^n r_{i,t}^2 \sum_{t=1}^n r_{j,t}^2}\right] \tag{23}$$

The expected cross-term above is zero since $\{e_t\}$ is independent of the $\{r_{i,t}\}$, which are functions of $\{x_t\}_{t=-p+1}^n$.

Using the fact that

$$\hat{\sigma}^2 \frac{\sum_{t=1}^n r_{i,t} r_{j,t}}{\sum_{t=1}^n r_{i,t}^2 \sum_{t=1}^n r_{j,t}^2} = \widehat{\text{cov}}(\hat{\beta}_i^c, \hat{\beta}_j^c) \tag{24}$$

where $\hat{\sigma}^2$ is the estimator of the error variance from a regression (with intercept) of y_t on $x_{t-1}, \ldots, x_{t-p}$ and v_t^c, the Lemma is proved following similar techniques used in the proof of Lemma 2 in Amihud and Hurvich (2004). □

Proof of Lemma 3: Let q be the residual vector in an OLS regression of v_t^c on x_{t-1}, $\cdots$, x_{t-p}. Since e is independent of $\{x_t\}_{t=-p+1}^n$, and since q is a function of $\{x_t\}_{t=-p+1}^n$, it follows that q is independent of e.

Next,

$$[\widehat{SE}(\hat{\phi}^c)]^2 = \frac{\hat{\sigma}^2}{\sum_{t=1}^n q_t^2} \quad .$$

Using the representation (18) together with the properties $\sum q_t v_t^c = \sum q_t^2$ and $\sum q_t x_{t-p} = \cdots = \sum q_t x_{t-1} = \sum q_t = \sum q_t(v_t - v_t^c) = 0$, we obtain

$$\hat{\phi}^c = \frac{\sum_{t=1}^n q_t y_t}{\sum_{t=1}^n q_t^2} = \phi + \frac{\sum_{t=1}^n q_t e_t}{\sum_{t=1}^n q_t^2} \quad . \tag{25}$$

Since $\{e_t\}$ is independent of $\{q_t\}$ and $E[e_t] = 0$, the expectation of the second term on the righthand side of the above equation is zero, and we obtain

$$\mathrm{var}[\hat{\phi}^c] = \sigma_e^2 E\left[\frac{1}{\sum_{t=1}^n q_t^2}\right] \quad . \tag{26}$$

Proceeding as in the proof of Lemma 2, we have

$$E\left[\frac{\hat{\sigma}^2}{\sum_{t=1}^n q_t^2} \mid X\right] = E\left[\frac{1}{n-p-2}\frac{e'(I-H)e}{\sum_{t=1}^n q_t^2} \mid X\right]$$

$$= \frac{1}{\sum_{t=1}^n q_t^2}\,\frac{1}{n-p-2}\,E[\sigma_e^2 \chi_{n-p-2}^2] = \sigma_e^2 \frac{1}{\sum_{t=1}^n q_t^2} \quad ,$$

where $X = [1_n, x_{t-1}, \ldots, x_{t-p}, v_t^c]$, $H = X(X'X)^{-1}X'$ and 1_n is an $n \times 1$ vector of ones.

Taking expectations of both sides and using the double expectation theorem yields

$$E\left[\frac{\hat{\sigma}^2}{\sum_{t=1}^n q_t^2}\right] = \sigma_e^2 E\left[\frac{1}{\sum_{t=1}^n q_t^2}\right] \quad .$$

The Lemma now follows from (26). $\square$

Bias of the OLS Estimators of the Autoregressive Coefficients: From Table 1 of Shaman and Stine (1988).

p	bias expressions for $\hat{\rho}_1, \ldots, \hat{\rho}_p$
1	$-\frac{1+3\rho_1}{n}$
2	$-\frac{1+\rho_1+\rho_2}{n}$, $-\frac{2+4\rho_2}{n}$
3	$-\frac{1+\rho_1+2\rho_3}{n}$, $-\frac{2-\rho_1+4\rho_2+\rho_3}{n}$, $-\frac{1+5\rho_3}{n}$
4	$-\frac{1+\rho_1+\rho_4}{n}$, $-\frac{2-\rho_1+2\rho_2+\rho_3+2\rho_4}{n}$, $-\frac{1-2\rho_1+5\rho_3+\rho_4}{n}$, $-\frac{2+6\rho_4}{n}$
5	$-\frac{1+\rho_1+2\rho_5}{n}$, $-\frac{2-\rho_1+2\rho_2+2\rho_4+\rho_5}{n}$, $-\frac{1-2\rho_1-\rho_2+5\rho_3+\rho_4+2\rho_5}{n}$, $-\frac{2-\rho_1+6\rho_4+\rho_5}{n}$, $-\frac{1+7\rho_5}{n}$

References

[1] Akaike, H., 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19, 716-723.

[2] Amihud, Y., and C. M. Hurvich, 2004. Predictive Regression: a Reduced-Bias Estimation Method. *Journal of Financial and Quantitative Analysis*, 39, 813–841.

[3] Amihud, Y., C. M. Hurvich, and Y. Wang, 2006. Hypothesis Testing in predictive regressions with AR(1) predictor. *Working paper*, NYU-Stern.

[4] Amihud, Y., C. M. Hurvich, and Y. Wang, 2008. Multiple-Predictor Regressions: Hypothesis Testing. *Review of Financial Studies*, forthcoming.

[5] Ang, A., and G. Bekaert, 2007, Stock Return Predictability: Is It There? *Review of Financial Studies*, 20, 651–707.

[6] Bernanke, B.S. and I. Mihov, 1998a, Measuring Monetary Policy. *Quarterly Journal of Economics*, 113, 869-902.

[7] Bernanke, B.S. and I. Mihov, 1998b, The Liquidity Effect and Long-Run Neutrality. *Carnegie-Rochester Conference Series on Public Policy*, 49, 149-194

[8] Bernanke, B.S. and A. Blinder, 1992, The Federal Funds Rate and the Channels of Monetary Transmission. *American Economic Review* 82, 901- 921.

[9] Brennan, M.J. and Y. Xia, 2005, Persistence, Predictability and Portfolio Planning. *Working paper*, UCLA and Wharton.

[10] Campbell, B., and J-M Dufour, 1997, Exact Nonparametric Tests of Orthogonality and Random Walk in the Presence of a Drift Parameter. *International Economic Review* 38, 151173.

[11] Campbell, J. Y., and M. Yogo, 2005, Efficient Tests of Stock Return Predictability, *Journal of Financial Economics*, 81, 27–60.

[12] Cavanagh, C.L., G. Elliott and J.H. Stock, 1995, Inference in Models with Nearly Integrated Regressors. *Econometric Theory*, 11, 1131-1147.

[13] Chen, W.W., and R.S. Deo, 2009, Bias Reduction and Likelihood Based Almost-Exactly Sized Hypothesis Testing in Predictive Regressions using the Restricted Likelihood. *Econometric Theory*, forthcoming.

[14] Eliasz, P. 2005, Optimal Median Unbiased Estimation of Coefficients on Highly Persistent Regressors. Princeton University.

[15] Estrella, A., and G. A. Hardouvelis, 1991, The Term Structure as a Predictor of Real Economic Activity. *Journal of Finance*, 46, 555–576.

[16] Fama, E.F., 1990, Stock Returns, Expected Returns, and Real Activity. *Journal of Finance*, 45, 1089–1108.

[17] Fama, E.F., and K.R. French, 1989, Business Conditions and Expected Returns on Stocks and Bonds. *Journal of Financial Economics*, 25, 23–49.

[18] Ferson, W. E., S. Sarkissian and T. T. Simon, 2003, Spurious Regressions in Financial Economics? *Journal of Finance* 58, 1393-1413.

[19] Granger, C. W. J., 1969, Testing for Causality and Feedback. *Econometrica*, 37, 424–438.

[20] Greene, W. H., 2002, *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.

[21] Hamilton, J. D., 1994, *Time Series Analysis*. Princeton, NJ: Princeton University Press.

[22] Hamilton, J. D., and D. H. Kim, 2002, A Reexamination of the Predictability of Economic Activity Using the Yield Spread. *Journal of Money, Credit and Finance*, 34, 340–360.

[23] Jansson, M. and M.J. Moreira, 2006, Optimal Inference in Regression Models with Nearly Integrated Regressors. *Econometrica* 74, 681714.

[24] Lint, C. R., and D. Stolin, 2003, The Predictive Power of the Yield Curve: A Theoretical Assessment. *Journal of Monetary Economics*, 50, 1603–1622.

[25] Lynch, A.W. and s. Tan, 2009. Multiple Risky Assets, Transaction Costs and Return Predictability: Allocation Rules & Implications for U.S. Investors. *Journal of Financial and Quantitative Analysis*, forthcoming.

[26] Maynard, A., and K. Shimotsu, 2007, Covariance-based Orthogonality Tests for Regressors with Unknown Persistence. *Econometric Theory*, to appear.

[27] Nicholls, D. F., and A. L. Pope, 1988, Bias in the Estimation of Multiple Autoregressions. *Australian Journal of Statistics*, 30A, 296–309.

[28] Plosser, C. I., and K. G. Rouwenhorst, 1994, International Term Structure and Real Economic Growth. *Journal of Monetary Economics*, 33, 133–155.

[29] Polk, C., S. Thompson, and T. Vuolteenaho, 2006, Cross-sectional Forecasts of the Equity Premium, *Journal of Financial Economics*, 81, 101–141.

[30] Santos, T. and P. Veronesi, 2006, "Labor Income and Predictable Stock Returns," *Review of Financial Studies*, 19, 1–44.

[31] Shaman, P., and R. A. Stine, 1988, The Bias of Autoregressive Coefficient Estimators. *Journal of the American Statistical Association*, 83, 842–848.

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

Table 1: Simulation Results of a AR(2) Predictive Regression Model

The model is,

$$y_t = \alpha + \beta_1 x_{t-1} + \beta_2 x_{t-2} + u_t \quad ,$$

$$x_t = \theta + \rho_1 x_{t-1} + \rho_2 x_{t-2} + v_t \quad ,$$

where the errors (u_t, v_t) are each serially independent and identically distributed as bivariate normal, with contemporaneous correlation,

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim_{iid} N(0, \Sigma) \quad , \quad \Sigma = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} \quad ,$$

It is known that we can write $u_t = \phi v_t + e_t$ and $\{e_t\}$ are IID and independent of $\{x_t\}$ and $\{v_t\}$.

In the simulation, $\beta_1 = \beta_2 = 0$, thus the estimated parameters in the table represent the bias. We set $\phi = -92.17$, $\sigma_e = 0.01844$ and $\sigma_v = 0.0007746$. These parameter values, as well as those of Case 1, are taken from the empirical estimation in Table 2.

For the autoregressive coefficients, there are two cases:
Case 1: $\rho_1 = 1.1053$ and $\rho_2 = -0.1430$; the roots of the AR(2) process are 0.9557 and 0.1496.
Case 2: $\rho_1 = 1.0553$ and $\rho_2 = -0.1430$; the roots of the AR(2) process are 0.8956 and 0.1597.

Panel A shows the parameter estimates and Panel B shows the realized size when the 5% nominal size is used for for both t-test—right-sided and two-sides (using standard t-values that correspond to 5% tail probability)—and the Wald test. The simulations are based on 10000 realizations.

Throughout, the ARM-estimated parameters are indicated by ^{c} and are boldfaced.

Panel A: Parameter Estimation

Case 1: $\rho_1 = 1.1053$, $\rho_2 = -0.1430$, with roots of 0.9557 and 0.1496

	n=200			n=50		
	Mean	Std Dev	RMSE	Mean	Std Dev	RMSE
$\hat{\beta}_1$	1.0302	6.3088	6.3923	5.2630	12.9280	13.9582
$\hat{\beta}_2$	0.6576	6.1324	6.1675	2.0304	11.4311	11.6106
$\hat{\beta}_1^c$	**0.1343**	**6.3117**	**6.3131**	**1.7929**	**12.9973**	**13.1203**
$\hat{\beta}_2^c$	**0.0127**	**6.2472**	**6.2472**	**-0.4394**	**12.2878**	**12.2957**
$\widehat{SE}(\hat{\beta}_1)$	6.3325	0.4732	–	11.9039	1.9964	–
$\widehat{SE}(\hat{\beta}_2)$	6.3277	0.4885	–	11.6697	2.2256	–
$\widehat{SE}(\hat{\beta}_1^c)$	**6.3352**	**0.4733**	–	**11.9468**	**1.9978**	–
$\widehat{SE}(\hat{\beta}_2^c)$	**6.4474**	**0.4977**	–	**12.5536**	**2.3953**	–
$\widehat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_2)$	-37.91	5.64	–	-120.46	43.58	–
$\widehat{\text{cov}}(\hat{\beta}_1^c, \hat{\beta}_2^c)$	**-38.62**	**5.74**	–	**-129.32**	**46.85**	–
True $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$		-36.02			-121.50	
True $\text{cov}(\hat{\beta}_1^c, \hat{\beta}_2^c)$		**-36.69**			**-130.42**	

Case 2: $\rho_1 = 1.0553$, $\rho_2 = -0.1430$, with roots of 0.8956 and 0.1597

	n=200			n=50		
	Mean	Std Dev	RMSE	Mean	Std Dev	RMSE
$\hat{\beta}_1$	0.9352	6.4778	6.5449	4.2003	13.4161	14.0582
$\hat{\beta}_2$	0.6623	6.3145	6.3491	2.6173	12.0492	12.3302
$\hat{\beta}_1^c$	**0.0619**	**6.4819**	**6.4822**	**0.8127**	**13.4897**	**13.5141**
$\hat{\beta}_2^c$	**0.0176**	**6.4328**	**6.2329**	**0.1905**	**12.9499**	**12.9513**
$\widehat{SE}(\hat{\beta}_1)$	6.5240	0.2737	–	12.6222	1.4211	–
$\widehat{SE}(\hat{\beta}_2)$	6.5229	0.2857	–	12.5416	1.5649	–
$\widehat{SE}(\hat{\beta}_1^c)$	**6.5280**	**0.2737**	–	**12.6715**	**1.4237**	–
$\widehat{SE}(\hat{\beta}_2^c)$	**6.6462**	**0.2912**	–	**13.4918**	**1.6851**	–
$\widehat{\mathrm{cov}}(\hat{\beta}_1, \hat{\beta}_2)$	-38.48	3.52	–	-132.85	33.52	–
$\widehat{\mathrm{cov}}(\hat{\beta}_1^c, \hat{\beta}_2^c)$	**-39.20**	**5.59**	–	**-142.48**	**36.08**	–
True $\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2)$		-36.42			-129.56	
True $\mathrm{cov}(\hat{\beta}_1^c, \hat{\beta}_2^c)$		**-37.08**			**-138.93**	

Panel B: Hypothesis Testing: Realized Sizes of Test with a Nominal Size of 5%

Case 1: $\rho_1 = 1.1053$, $\rho_2 = -0.1430$, with roots of 0.9557 and 0.1496

	t-test				Wald test	
	n=200		n=50		n=200	n=50
	Right-tailed	Two-tailed	Right-tailed	Two-tailed		
$\hat{\beta}_1$	6.7%	5.4%	12.5%	9.5%	7.4%	12.7%
$\hat{\beta}_2$	5.6%	4.5%	7.3%	5.7%	7.4%	12.7%
$\hat{\beta}_1^c$	**5.2%**	**5.0%**	**8.2%**	**7.7%**	**7.0%**	**9.9%**
$\hat{\beta}_2^c$	**4.5%**	**4.6%**	**4.9%**	**5.1%**	**7.0%**	**9.9%**

Case 2: $\rho_1 = 1.0553$, $\rho_2 = -0.1430$, with roots of 0.8956 and 0.1597

	t-test				Wald test	
	n=200		n=50		n=200	n=50
	Right-tailed	Two-tailed	Right-tailed	Two-tailed		
$\hat{\beta}_1$	6.5%	5.1%	10.2%	8.0%	6.3%	9.5%
$\hat{\beta}_2$	5.5%	4.5%	7.2%	5.4%	6.3%	9.5%
$\hat{\beta}_1^c$	**5.1%**	**4.9%**	**6.8%**	**6.9%**	**5.8%**	**8.1%**
$\hat{\beta}_2^c$	**4.5%**	**4.5%**	**5.1%**	**4.7%**	**5.8%**	**8.1%**

Table 2: Market Return Predicted by Lagged Dividend Yield

$VWNY$ is the value-weighted NYSE quarterly stocks return and DY is dividend yield on these stocks for the end of the quarter. The estimation period is 1946-1994. The estimated model is:

$$\begin{aligned} VWNY_t &= \alpha + \beta_1 \log(DY_{t-1}) + \beta_2 \log(DY_{t-2}) + u_t \\ \log(DY_t) &= \theta + \rho_1 \log(DY_{t-1}) + \rho_2 \log(DY_{t-2}) + v_t \end{aligned}$$

We present results for OLS regressions and for the corresponding Augmented Regression Method (ARM). For each parameter, we report the point estimate as well as the corresponding t-statistic (in parenthesis). The hypothesis testing are two-sided. * and ** indicate significance at 5% and 1% levels, respectively. The joint Wald test is a test of the joint hypothesis that the vector (β_1, β_2) is (0,0).

Model	Est. Method	$\hat{\beta}_1$ or $\hat{\beta}_1^c$	$\hat{\beta}_2$ or $\hat{\beta}_2^c$	$\hat{\rho}_1^c$	$\hat{\rho}_2^c$	$\hat{\phi}^c$	Joint Wald Test
A	OLS	-2.9696	10.8244				14.35**
Predictor		(-0.44)	(1.58)				p-value: 0.0008
is AR(2)	ARM(2)	-3.8937	10.1597	1.1053	-0.1430	-92.17	3.91
		(-1.45)	(1.46)	(49.31)**	(-1.96)*	(-53.50)**	p-value: 0.142
B	OLS	7.2814					
Assume		(3.43)**					
predictor	ARM(1)	5.4569		0.9729		-91.72	
is AR(1)		(2.53)*		(42.73)**		(-36.10)**	

Table 3: Simulation Results for Multivariate $ARM(p)$

Bivariate Predictor

Predictors are $AR(2)$, each having $\rho_1 = 1.1053$, $\rho_2 = -0.1430$, with roots of 0.9557 and 0.1496

Tests have Nominal Size 0.05

t-test Results for (Right-Tailed, Two-Tailed)

	OLS			ARM		
Parameter	Mean	Std Dev	t-test Size	Mean	Std Dev	t-test Size
($n = 50$)						
$\beta_{1,1}$	8.81	29.07	0.109, 0.082	1.95	14.97	0.088, 0.074
$\beta_{2,1}$	1.04	28.39	0.067, 0.064	–0.420	14.95	0.053, 0.054
$\beta_{1,2}$	8.69	28.38	0.097, 0.075	2.04	14.49	0.070, 0.067
$\beta_{2,2}$	1.08	28.01	0.053, 0.059	–0.693	14.43	0.042, 0.046
Wald Test Size		0.166			0.115	
($n = 200$)						
$\beta_{1,1}$	1.77	13.22	0.069, 0.055	0.21	6.86	0.056, 0.052
$\beta_{2,1}$	0.42	13.25	0.050, 0.045	–0.005	6.83	0.048, 0.051
$\beta_{1,2}$	1.40	13.63	0.063, 0.061	0.004	6.92	0.056, 0.052
$\beta_{2,2}$	0.65	13.60	0.062, 0.049	0.084	6.94	0.050, 0.047
Wald Test Size		0.092			0.081	