TECHNICAL WORKING PAPER SERIES

PREDICTIVE REGRESSIONS

Robert F. Stambaugh

Technical Working Paper 240
http://www.nber.org/papers/T0240

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 1999

Comments by an anonymous referee, Doron Avramov, John Campbell, Ľuboš Pástor, and workshop participants at the University of Chicago and the University of Pennsylvania are gratefully acknowledged. The author also acknowledges the support provided by his appointment during the 1997-98 academic year as a Marvin Bower Fellow at Harvard Business School, where portions of this research were conducted. This study includes results from the author's 1996 working paper, "Bias in Regressions with Lagged Stochastic Regressors." Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

© 1999 by Robert F. Stambaugh. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Predictive Regressions
Robert F. Stambaugh
NBER Technical Working Paper No. 240
May 1999
JEL No. C32, C11, G11

ABSTRACT

When a rate of return is regressed on a lagged stochastic regressor, such as a dividend yield, the regression disturbance is correlated with the regressor's innovation. The OLS estimator's finite-sample properties, derived here, can depart substantially from the standard regression setting. Bayesian posterior distributions for the regression parameters are obtained under specifications that differ with respect to (i) prior beliefs about the autocorrelation of the regressor and (ii) whether the initial observation of the regressor is specified as fixed or stochastic. The posteriors differ across such specifications, and asset allocations in the presence of estimation risk exhibit sensitivity to those differences.

Robert Stambaugh
Finance Department
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6367
and NBER
stambaugh@wharton.upenn.edu

1. Introduction

Many empirical studies in economics and finance investigate regressions of the form

$$y_t = \alpha + \beta x_{t-1} + u_t, \tag{1}$$

where y_t reflects a change in an asset's price during period t, x_{t-1} is a lagged variable related to asset prices at the end of period $t-1$, and u_t is the regression's disturbance. Examples of such a regression occur when y_t is the return on a portfolio of common stocks, and x_{t-1} is a dividend yield or a function of current or lagged interest rates.[1] A regression as in (1) also arises in studies of fixed-income markets, where y_t is the excess return on a bond portfolio or a change in an interest rate, and x_{t-1} is an interest rate, yield spread, or forward rate.[2] Investigations of the foreign-exchange market often include a regression as in (1), where y_t is the change in the spot rate of exchange, and x_{t-1} is the spread between the forward and spot exchange rates.[3]

A standard regression-model assumption maintained here is that u_t is serially uncorrelated and has zero expectation conditional on $\{x_{t-1}, x_{t-2}, \ldots\}$. An assumption that typically fails to hold in the examples noted above is that u_t has zero expectation conditional on $\{x_s, \text{ for all } s\}$, and this is the assumption used to obtain finite-sample results in the standard setting. In particular, if x_{t-1} depends on asset prices at the end of period $t-1$, then the value of that regressor at the end of period t reflects changes in asset prices during period t, as does y_t, so $\mathrm{E}\{u_t | x_t, x_{t-1}\} \neq 0$. More generally,

$$\mathrm{E}\{u_t | x_s, x_w\} \neq 0, \quad s < t \leq w, \tag{2}$$

since a price change during period t is correlated with the change in the regressor over an interval that includes period t.

A consequence of (2) is that finite-sample estimation and inference become less straightforward, for at least two reasons. First, the ordinary least squares (OLS) estimators of the coefficients in (1) are biased and have sampling distributions that differ from those in the standard setting, and a classical or "frequentist" approach must account for such depar-

[1] There are many examples, including Fama and Schwert (1977), Rozeff (1984), Keim and Stambaugh (1986), Campbell (1987), and Fama and French (1988). See Kothari and Shanken (1997) and Pontiff and Schall (1998) for recent examples.

[2] A few examples include Shiller, Campbell, and Schoenholtz (1983), Fama (1984a), Keim and Stambaugh (1986), and Fama and Bliss (1987).

[3] Early examples include Bilson (1981) and Fama (1984b).

tures.[4] Second, differences between classical and Bayesian methods become more apparent in the presence of (2), whereas those approaches are distinguished less often in the standard regression setting. In the standard setting, classical confidence intervals correspond to Bayesian highest-posterior-density regions under diffuse priors, and the p-value for a positive one-tailed test of $\beta = 0$ is identical to the posterior probability that $\beta \leq 0$ (see Box and Tiao, 1973). That correspondence no longer obtains in the current setting, wherein a Bayesian could, for example, assign low probability to $\beta \leq 0$ at the same time the frequentist accepts that hypothesis because its associated p-value is large. Such an example is provided in this study, which addresses both classical and Bayesian issues.

The example chosen for illustration is one in which y_t is the return on the aggregate stock-market portfolio and x_{t-1} is that portfolio's dividend yield. Such a regression has received substantial attention in the finance literature, but an additional motivation for selecting this example highlights another distinction sometimes made in contrasting classical and Bayesian approaches: data description versus decision making. On one hand, a classical p-value or confidence region conveys information about the data in an objective fashion, and one might argue that the dependence on prior beliefs makes Bayesian analysis less effective in communicating a description of the data (e.g., Stock, 1991). On the other hand, one might argue that reporting implications for decisions describes the data in a more relevant manner, and a Bayesian framework is better suited to that purpose. Kandel and Stambaugh (1996), for example, use a Bayesian framework to explore the implications for a stocks-versus-cash allocation associated with a regression as in (1), where y_t is the excess stock return. They find that such a characterization of the data often communicates a different message than that delivered by p-values for the hypothesis $\beta = 0$. The regression of stock return on dividend yield affords an exploration of the study's Bayesian methods in an asset-allocation context.

The paper proceeds as follows. Sections 2 through 4 underscore the finite-sample nature of the regression problem along several dimensions. In Section 2, the finite-sample distribution and moments of the OLS estimator of β are derived analytically and computed for the regression of excess return on dividend yield. The exact moments and p-values can exhibit large differences from their counterparts in the standard regression setting. For example, in the overall 70-year period from 1927–96, the bias equals one-third of the OLS estimate

[4]Early demonstrations of this point include Mankiw and Shapiro (1986) and Stambaugh (1986). Monte Carlo or bootstrap simulations have been used for finite-sample inference in this problem by a number of studies, including Nelson and Kim (1993), in an investigation of stock-return predictability, Bekaert, Hodrick, and Marshall (1997), in an investigation of the expectations hypothesis of the term structure, and Mark (1995), in an investigation of exchange-rate predictability.

for that period, and the correct p-value for the hypothesis $\beta = 0$ is roughly three times the value based on the usual t-statistic.

Section 3 analyzes Bayesian posterior distributions for the regression coefficients and finds that those distributions exhibit sensitivity to what some might view as minor differences in the prior or the likelihood function. For example, treating the initial observation x_0 as stochastic and drawn from the regressor's stationary distribution, as opposed to treating x_0 as a fixed value, can produce a substantial difference in the posterior mean of β and in the maximum-likelihood estimate, even in a 45-year sample. The posterior distribution of β is also sensitive to specification of the prior, even when the different specifications are all intended to represent "noninformative" beliefs.

Section 4 considers an asset-allocation problem for an investor whose perceived distribution of future returns is given by the predictive distribution arising from one of the Bayesian specifications analyzed in Section 3. For both short and long investment horizons, the optimal stock allocation of a buy-and-hold investor exhibits sensitivity to the alternative specifications of the prior and the likelihood. Also observed is the possibility that, at long horizons, the investor might actually allocate more to stocks at lower levels of the current dividend yield (lower expected returns). That behavior arises due to conditional skewness in the predictive distribution of long-horizon returns. The skewness can be traced to effects of finite-sample parameter uncertainty or "estimation risk," particularly uncertainty about the regressor's persistence.

The analyses in Sections 2 through 4 focus on settings in which a single independent variable appears on the right-hand side of the predictive regression in (1). This simplest setting proves useful in developing analytical results as well as insights, but much of the methodology can be extended to a setting with multiple predictive variables, as discussed in Section 5. Section 6 reviews the conclusions.

2. Ordinary least squares in finite samples

It is assumed throughout that x_t obeys a first-order autoregressive (AR(1)) process,

$$x_t = \theta + \rho x_{t-1} + v_t. \tag{3}$$

The vector $(u_t\ v_t)'$ is assumed to be normally distributed, independently across t, with mean zero and covariance matrix

$$\text{cov}\left\{\begin{pmatrix} u_t \\ v_t \end{pmatrix}, (u_t\ v_t)\right\} \equiv \Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}. \tag{4}$$

This distributional assumption permits exact finite-sample results, both classical and Bayesian.[5] In this section, it is also assumed that $|\rho| < 1$. The latter assumption implies stationarity of the regressor, although, as in the regression of return on dividend yield, the value of ρ can be close to 1. The OLS estimators of the coefficients in (1) are given by

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = (X'X)^{-1}X'y, \tag{5}$$

where $y \equiv (y_1\ \ldots\ y_T)'$, $X \equiv [\iota_T\ x_{(\ell)}]$, $x_{(\ell)} \equiv (x_0\ \ldots\ x_{T-1})'$, and ι_T denotes a $T \times 1$ vector of 1's.

Before proceeding to a more formal treatment of finite-sample properties, it may be useful to understand how $\hat{\beta}$ is biased under the simplest possible setting in which the estimator is defined. That is, consider repeated samples of only two observations, (x_0, y_1) and (x_1, y_2), so $\hat{\beta}$ in each sample is simply the slope of the line connecting those points. For this purpose let $\beta = 0$, $\rho \approx 1$, and $\sigma_{uv} < 0$. First consider the samples in which $x_1 > x_0$, or essentially $v_1 > 0$ (since $\rho \approx 1$). On average across such samples, $y_2 = E\{y_t\}$ (since $\beta = 0$), $y_1 < E\{y_t\}$ (since $\sigma_{uv} < 0$ implies u_1 is on average negative when $v_1 > 0$), and therefore $\hat{\beta}$ is positive (since $y_2 > y_1$ and $x_1 > x_0$). On average across the samples in which $x_1 < x_0$, or $v_1 < 0$, $y_2 = E\{y_t\}$ as before, but now $y_1 > E\{y_t\}$, so again $\hat{\beta}$ is positive ($y_2 < y_1$ and $x_1 < x_0$). Thus, on average across all samples, $\hat{\beta} > 0$, i.e., $\hat{\beta}$ is upward biased. Note that if $\sigma_{uv} > 0$, the same analysis leads to a downward bias in $\hat{\beta}$. Note also that the bias disappears as σ_{uv} approaches zero, since the sign of v_1 then has no association with that of u_1. Finally, note that the bias shrinks as ρ approaches zero, since the sign of $x_1 - x_0$ is then linked less tightly to the sign of v_1 and, thereby, to the sign of $y_2 - y_1$ (although even with $\rho = 0$ there is still some association and hence some bias). As shown below, σ_{uv} and ρ play similar roles in a more general setting with T observations. Of course, as T increases, the scatter of points essentially becomes a horizontal cloud of these two-point clusters (with $\beta = 0$), and the bias in the fitted slope approaches zero.

The finite-sample properties of $\hat{\beta}$ can be derived by first recognizing that the estimator can be represented as a ratio of quadratic forms:

[5] Asymptotic approaches to inferences about β are developed under weaker distributional assumptions in Elliott and Stock (1994) and Cavanagh, Elliott, and Stock (1995), where ρ is entertained as "local to unity" in the sense that it approaches 1 as the sample size grows.

Proposition 1. *The finite-sample distribution of* $\hat{\beta} - \beta$ *depends on* ρ *and* Σ *but not on* α, β, *or* θ, *and*

$$\hat{\beta} - \beta = \frac{w'Aw}{w'Bw}\,, \tag{6}$$

where $w = (u' \;\; x'_{(\ell)} - \mu_x \iota'_T)'$, $u = (u_1, \ldots, u_T)'$, $\mu_x = E\{x_t\}$, $E\{w\} = 0$,

$$cov\{w, w'\} = \Omega = LL' = \begin{bmatrix} \sigma_u^2 I_T & \sigma_{uv} G \\ \sigma_{uv} G' & \sigma_v^2 H \end{bmatrix}, \tag{7}$$

G *is a* $T \times T$ *matrix whose* (i, j) *element is* ρ^{j-i-1} *for* $i < j$ *and zero otherwise,* H *is a* $T \times T$ *matrix whose* (i, j) *element is* $[1/(1-\rho^2)]\rho^{|i-j|}$,

$$A = \frac{1}{2}\begin{bmatrix} 0 & M \\ M & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & M \end{bmatrix}, \tag{8}$$

$M = I_T - (1/T)\iota_T \iota'_T$, *and* I_T *denotes the* $T \times T$ *identity matrix.*

Proof: see the Appendix.

The representation of $\hat{\beta} - \beta$ in (6) allows the distribution and moments of $\hat{\beta}$ to be derived analytically using results from the literature on quadratic forms. The cumulative distribution of $\hat{\beta}$, given by the following proposition, relies on a result by Imhof (1961).

Proposition 2. *For any fixed* β_0,

$$Prob\{\hat{\beta} > \beta_0\} = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty q^{-1} \prod_{i=1}^{M} (1 + \psi_i^2 q^2)^{-n_i/4} \sin\left(\frac{1}{2}\sum_{i=1}^{M} n_i \tan^{-1}(\psi_i q)\right) dq, \tag{9}$$

where $\psi_i, i = 1, \ldots, M$, *denote the* M *distinct nonzero eigenvalues of* $L'[A - (\beta_0 - \beta)B]L$, *and* n_i *is the multiplicity of* ψ_i.

Proof: see the Appendix.

The finite-sample moments of $\hat{\beta}$, given in the following proposition, are obtained by applying a result from Magnus (1986) to the representation given in (6).

Proposition 3. *For each integer* s, $1 \leq s < (T-1)$,

$$m'_s \equiv E\{(\hat{\beta} - \beta)^s\} = s2^s \sum_i \gamma_s(\nu_i) \int_0^\infty q^{s-1} |\Delta| \prod_{j=1}^{s} (\mathrm{tr}\, R^j)^{n_{ij}} dq, \tag{10}$$

where the summation is over all vectors $\nu_i = (n_{i1}, n_{i2}, \ldots, n_{is})$ *whose* s *elements are non-negative integers satisfying* $\sum_{j=1}^{s} j n_{ij} = s$,

$$\gamma_s(\nu_i) = \prod_{j=1}^{s} [n_{ij}!(2j)^{n_{ij}}]^{-1}\,, \tag{11}$$

and where the $2T \times 2T$ matrices Δ and R are constructed as follows. Let P be a $2T \times 2T$ matrix such that $P'P = I_{2T}$ and $P'L'BLP = \Lambda$, a diagonal matrix. Then $\Delta = (I_{2T}+2q\Lambda)^{-1/2}$ and $R = \Delta P'L'ALP\Delta$.

Proof: see the Appendix.

The moments in Proposition 3 are noncentral, since $\mathrm{E}\{\hat{\beta}\} \neq \beta$, but the central moments are easily obtained using standard relations between central and noncentral moments (e.g., Kendall and Stuart, 1977, p. 58):

Corollary. *Let m_s denote the central moment $E\{(\hat{\beta} - m_1)^s\}$, where $m_1 \equiv E\{\hat{\beta}\} = m'_1 + \beta$. For $1 < s < (T-1)$,*

$$m_s = \sum_{j=0}^{s} \binom{s}{j} m'_{s-j}(-m'_1)^j, \tag{12}$$

and, in particular,

$$m_2 = m'_2 - m'^2_1, \tag{13}$$

$$m_3 = m'_3 - 3m'_1m'_2 + 2m'^3_1, \quad \textit{and} \tag{14}$$

$$m_4 = m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1. \tag{15}$$

Table 1 reports finite-sample properties of $\hat{\beta}$, under the normality assumption, for a regression in which y_t is the continuously compounded excess return during month t on the value-weighted portfolio of NYSE stocks, and x_{t-1} is that portfolio's dividend yield, defined as dividends paid during months $t-12$ through $t-1$ divided by the portfolio's value at the end of month $t-1$. The portfolio's "excess" return is its rate of return minus the rate on a one-month Treasury bill, where both returns are continuously compounded. Results are shown across four sample periods. Part A of Table 1 reports the finite-sample bias, standard deviation, skewness ($m_3/m_2^{3/2}$), and kurtosis (m_4/m_2^2) of $\hat{\beta}$, as well as the p-value for a test of $\beta = 0$ versus $\beta > 0$. The moments are computed using Proposition 3, and the "true" p-value is computed as the probability in Proposition 2 with β set to zero and β_0 set equal to the sample value of $\hat{\beta}$.[6] Computing the quantities in Part A requires the true (unknown) values of ρ and Σ. For each sample period, ρ is set equal to that period's least-squares estimate from equation (3), and Σ is set equal to the sample covariance matrix of the least-squares residuals from (1) and (3). Those values for ρ and Σ, as well as the sample size T and the realized sample value of $\hat{\beta}$, are given in Part C of Table 1. Part B of the table reports

[6]The required integrals are computed using standard numerical integration methods.

the corresponding moments and p-values implied by the standard regression model. The standard deviations in Part B depend on σ_u^2 and are conditioned on the sample values of x_{t-1}, which are assumed to be held fixed in repeated samples in the standard setting.

The results in Table 1 reveal marked differences between the true finite-sample properties of $\hat{\beta}$ and those implied by the standard setting. In this application, $\hat{\beta}$ is biased upward, positively skewed, and has higher variance and kurtosis than the (normal) sampling distribution of the OLS estimator in the standard setting. Even for the overall 70-year period ($T = 840$), the bias (0.07) is about one-third of the OLS estimate (0.21), and the skewness and kurtosis are 0.7 and 3.8. For the shortest and most recent period, still twenty years long ($T = 240$), $\hat{\beta}$ has a bias (0.42) nearly as large as its standard deviation (0.45) and more than twice its realized value (0.19). When computed using the t-statistic for the standard regression model, the one-tailed p-values for the hypothesis $\beta = 0$ are equal to 0.06 for the overall 70 year period and 0.02 for the 45-year period from 1952-96, whereas the true p-values for those periods are equal to 0.17 and 0.15.

The bias in $\hat{\beta}$ is related to the bias in $\hat{\rho}$, the sample first-order autocorrelation of x_t.[7] Define

$$\begin{bmatrix} \hat{\theta} \\ \hat{\rho} \end{bmatrix} = (X'X)^{-1}X'x, \tag{16}$$

where $x = (x_1 \ \ldots \ x_T)'$.

Proposition 4.

$$E\{\hat{\beta} - \beta\} = \frac{\sigma_{uv}}{\sigma_v^2} E\{\hat{\rho} - \rho\} \tag{17}$$

Proof: see the Appendix.

The bias in $\hat{\rho}$ is negative, and since price appears in the denominator of dividend yield, the unexpected return, u_t, is negatively correlated with the innovation in dividend yield, v_t. In the regressions of return on dividend yield used to construct Table 1, the value of σ_{uv}/σ_v^2 ranges between -13.6 and -22.3 across the four sample periods. Thus, from (17), the magnitude of the positive bias in $\hat{\beta}$ is many times that of the negative bias in $\hat{\rho}$. At the same time, β can be of the same or smaller magnitude as ρ: the values of $\hat{\beta}$ in Table 1 are all less than 0.5, whereas the values of $\hat{\rho}$ range between 0.94 and 0.99. As a result, the bias in $\hat{\rho}$ can be only a small fraction of ρ, but the bias in $\hat{\beta}$ can be a substantial fraction of β. Exact first and second finite-sample moments of $\hat{\rho}$, when $|\rho| < 1$ and v_t is normal,

[7] The results in (17) and (18) appear in Stambaugh (1986).

are derived and analyzed by Sawa (1978) and Nankervis and Savin (1988). The latter study reports, for example, that when $T = 200$ and $\rho = 0.99$, the bias in $\hat{\rho}$ is equal to -0.024, or only about 2.4% of ρ. With $\sigma_{uv}/\sigma_v^2 = -15$, equation (17) gives the corresponding bias in $\hat{\beta}$ as 0.36, which can be a substantial fraction of β. For both $\hat{\rho}$ and $\hat{\beta}$ in this example, the standard deviations of the OLS estimators are of similar magnitudes to their biases, so the biases in $\hat{\rho}$ and $\hat{\beta}$ are more comparable when viewed in that sense.

Under the normality assumption, a well-known approximation for the bias in $\hat{\rho}$, to order $1/T$, is given by $-(1 + 3\rho)/T$, as shown by Marriott and Pope (1954) and Kendall (1954). Thus, Proposition 4 yields a similar approximation for the bias in $\hat{\beta}$:

Corollary.

$$E\{\hat{\beta} - \beta\} = -\frac{\sigma_{uv}}{\sigma_v^2}\left(\frac{1 + 3\rho}{T}\right) + O(1/T^2). \tag{18}$$

The error in the approximation in (18) can be nontrivial, even for values of T that seem large for many purposes. In the regression of return on dividend yield, the true bias in $\hat{\beta}$ is equal to 0.42 for the 1977–96 period ($T = 240$), as reported in Table 1, whereas (18) gives a value of 0.35, which understates the bias by roughly 16%. The relative error in the approximation is decreasing in T, and (18) understates the true bias by about 4% for the example based on the 1927–96 period ($T = 840$).

As explained previously, the exact finite-sample moments and p-values in Table 1 depend on ρ and Σ. The true values of those parameters are unknown in practice, so in any given application one cannot know precisely the exact finite-sample moments of $\hat{\beta}$. The finite-sample properties in Table 1 are computed using the values of ρ and Σ obtained in the OLS estimation. Many of those computations are relatively insensitive to small changes in the parameters. For example, if the value of ρ is increased from $\hat{\rho}$ to $\hat{\rho} + (1 + 3\hat{\rho})/T$, a bias correction of order $(1/T)$, the p-values and the biases of $\hat{\beta}$ reported for the first three sample periods in Part A of Table 1 are changed by no more than 0.01. The standard deviations decline slightly, by 10% or less, whereas the skewness and kurtosis both increase, typically by around 10%.

In the fourth subperiod, increasing $\hat{\rho}$ by the bias in $\hat{\rho}$ (conditional on $\rho = \hat{\rho}$) produces a value greater than 1, so that bias-adjusted estimate cannot be used as a value of ρ in applying Propositions 2 and 3. Such an outcome illustrates a potentially unappealing aspect of estimating β and ρ by applying bias corrections. Suppose one assumes $|\rho| < 1$ and estimates ρ by adding the bias correction to the OLS estimator $\hat{\rho}$. This procedure can

produce a value greater than 1, as illustrated here, and one might be reluctant to accept such an estimate as a sensible value of ρ, even with the knowledge that this bias-corrected estimator would have the correct average across hypothetical repeated samples. Given the link in (17) between the biases in $\hat{\rho}$ and $\hat{\beta}$, applying the corresponding bias correction to $\hat{\beta}$ might then also be unappealing. Alternative approaches for obtaining estimates of β and ρ from the sample at hand are pursued in the next section.

3. Bayesian approaches

Finite-sample inferences about the parameters in (1) can also be pursued in a Bayesian setting. The results of the previous section indicate that, based on correctly computed p-values, the hypothesis that dividend yields fail to predict monthly stock returns would not be rejected at conventional significance levels. As mentioned earlier, in the standard Bayesian regression model with diffuse priors, the one-tailed p-value for the hypothesis $\beta = 0$ is identical to the posterior probability that $\beta \leq 0$. In the presence of (2), that finite-sample equivalence between p-values and posterior tail probabilities no longer obtains. In the standard setting, β is the mean of the sampling distribution of $\hat{\beta}$, and $\hat{\beta}$ is the mean of the posterior distribution of β. In the current setting, β is no longer the sampling mean of $\hat{\beta}$, as discussed in the previous section, although $\hat{\beta}$ is still the posterior mean of β for a particular specification of the prior and likelihood, as will be explained below. In general, however, the estimates and inferences delivered by a Bayesian approach to the regression problem considered here depart from their classical counterparts.

3.1. Methodology

Let $b \equiv (\alpha\ \beta\ \theta\ \rho)'$. A posterior density for b and Σ is computed as

$$p(b, \Sigma|D) \propto p(b, \Sigma)L(b, \Sigma; D), \tag{19}$$

where $p(b, \Sigma)$ denotes the prior density, L denotes the likelihood function, and D denotes the available data, which consist of $z \equiv (y'\ x')'$ and the initial observation of the regressor, x_0. The marginal posterior $p(\beta|D)$ is obtained by integrating (19) with respect to Σ and the other elements of b. The mean of the posterior density is commonly proposed as an estimator in a Bayesian setting, and values of $\mathrm{E}\{\beta|D\}$ are reported here for several alternative

specifications of the prior and the likelihood.[8] In addition, the posterior density yields probabilities for composite hypotheses, such as prob$\{\beta \leq 0\}$, and, as will be observed, the inferences associated with such probabilities can contrast with those provided by frequentist p-values.

Recall that the disturbance vector $(u_t \; v_t)'$ is assumed to obey a bivariate normal distribution. It is well known that the OLS estimators in (5) and (16) are then also maximum-likelihood estimators (MLE's) when the initial observation of the regressor, x_0, is assumed to be nonstochastic. The likelihood function under the latter assumption, the "conditional" likelihood, is given by

$$L_c(b, \Sigma; D) = p(z|x_0, b, \Sigma) = (2\pi|\Sigma|)^{-(T/2)} \exp\left\{-\frac{1}{2}(z - Zb)'(\Sigma^{-1} \otimes I_T)(z - Zb)\right\}, \quad (20)$$

where $Z = I_2 \otimes X$, and (20) is maximized at

$$\hat{b} \equiv (\hat{\alpha} \; \hat{\beta} \; \hat{\theta} \; \hat{\rho})' = (Z'Z)^{-1}Z'z. \quad (21)$$

As explained below, $\hat{b}$ is also the posterior mean of b when the likelihood function is given by (20) and $p(b, \Sigma)$ follows the standard specification for a noninformative prior in a multivariate regression model.

A common approach to specifying a noninformative or "diffuse" prior follows from Jeffreys (1961). If δ denotes a vector containing the unknown parameters, then an application of Jeffreys's invariance arguments leads to the specification

$$p(\delta) \propto \left|-\mathrm{E}\left\{\frac{\partial^2 \log L(\delta; D)}{\partial\delta\partial\delta'}\right\}\right|^{1/2}, \quad (22)$$

where the expectation is with respect to $p(D|\delta)$. The likelihood function in (20) also arises in the standard multivariate regression model, wherein Z is essentially viewed as nonstochastic. In that model, the prior is derived under the assumption $p(b, \Sigma) = p(b)p(\Sigma)$, and (22) is then applied separately for b and Σ. That procedure leads to the diffuse prior

$$p(b, \Sigma) \propto |\Sigma|^{-3/2}. \quad (23)$$

If the prior in (23) is combined with the conditional likelihood function in (20), then the resulting posterior density for b, a matrix t distribution, is given by standard results for the Bayesian multivariate regression model.[9] That posterior has the property that $\mathrm{E}\{\beta|D\} = \hat{\beta}$,

[8]The posterior mean has minimum posterior expected loss under a squared-error loss function (see Berger, 1985).

[9]For a Bayesian analysis of the standard multivariate regression model, including a discussion of the Jeffreys prior and the resulting posterior densities, see Zellner (1971, pp. 41–53 and pp. 224–233.)

even though, as discussed in the previous section, $\mathrm{E}\{\hat{\beta}\} \neq \beta$ (where $p(D|b, \Sigma)$ is used to take the latter expectation).

Although $\hat{b}$ emerges as the posterior mean of b with the likelihood in (20) and the prior in (23), that specification has several characteristics to which some might object. The likelihood function in (20) is subject to the criticism that treating the initial observation x_0 as non-stochastic can be inappropriate. If x_0 is non-stochastic, then that observation provides essentially no information about the unknown parameters of the model, but additional information can be provided by x_0 if it is instead a realization of the same stochastic process generating $x_1, \ldots, x_T$. The latter scenario seems more likely in finance and economics, where x_t is often a dividend yield, interest rate, or similar economic variable. If, for example, $|\rho| < 1$ and the process for x_t has run for a substantial time prior to the sample period, then x_0 is a realization of a normal variate with mean $\theta/(1-\rho)$ and variance $\sigma_v^2/(1-\rho^2)$, so x_0 provides information about θ, ρ, and σ_v. In essence, if x_0 is stochastic, then $p(b, \Sigma|x_0)$ can differ from $p(b, \Sigma)$, so using the latter prior with the conditional likelihood in (20) can be inappropriate. When it is assumed that $|\rho| < 1$, the density of x_0 given b and Σ is given by

$$p(x_0|b, \Sigma) = \left(\frac{1-\rho^2}{2\pi\sigma_v^2}\right)^{1/2} \exp\left\{-\frac{1-\rho^2}{2\sigma_v^2}\left(x_0 - \frac{\theta}{1-\rho}\right)^2\right\}. \tag{24}$$

The resulting "exact" likelihood function, which reflects the stochastic nature of x_0, is

$$L_e(b, \Sigma; D) = p(z, x_0|b, \Sigma) = p(z|x_0, b, \Sigma)p(x_0|b, \Sigma), \tag{25}$$

where $p(z|x_0, b, \Sigma)$ is given in (20).[10]

A possible objection to the prior in (23) is that non-stationary processes for x_t are entertained, i.e. nonzero prior probability is assigned to $|\rho| \geq 1$. Stationarity of the predictive variable is a property that one might wish to impose a priori in many applications. In (23), the implied prior density on ρ is "flat," i.e., $p(\rho)d\rho \propto d\rho$, so each fixed-length interval for ρ is assigned equal prior mass. A flat prior is one specification for noninformative beliefs about ρ, and the analysis below considers an alternative to (23) that preserves a flat marginal prior on ρ but simply confines that parameter to the stationary region, i.e., $p(\rho) = 1/2,\ \rho \in (-1,\ 1)$. If the marginal priors on the remaining parameters remain as in (23), then the joint prior is simply restated as

$$p(b, \Sigma) \propto |\Sigma|^{-3/2}, \quad \rho \in (-1,\ 1). \tag{26}$$

[10]For moving-average and autoregressive processes, Box and Jenkins (1970) derive "exact" likelihood functions that incorporate the stochastic nature of the initial observations.

The priors in both (23) and (26) are flat with respect to ρ. The issue of flat versus non-flat priors has received substantial attention in the context of the AR(1) model in (3). As Sims (1988) and Sims and Uhlig (1991) observe, conditional on x_0, a flat prior for ρ and a normal likelihood imply a posterior for ρ that is symmetric around $\hat{\rho}$, whereas the sampling distribution of $\hat{\rho}$ is not symmetric around ρ. Sims and and Uhlig (1991) use such a framework to demonstrate contrasts between Bayesian posterior tail probabilities and frequentist p-values. Phillips (1991) argues that a flat prior for ρ does not appropriately represent ignorance and suggests a Jeffreys prior be used instead.[11] Box and Jenkins (1970) also suggest the use of Jeffreys priors in Bayesian estimation of time series models. Citing earlier work by Perks (1947) and Welch and Peers (1963), Phillips notes that one characterization of a Jeffreys prior as representing "ignorance" is that it assigns higher density to regions of the parameter space where asymptotic confidence regions have lower anticipated volume. These priors also possess a well known invariance property, as noted by Jeffreys (1961). That is, if an alternative set of parameters is obtained as a one-to-one transformation of the original set, a Jeffreys prior on the alternative set results in a posterior density that is equivalent, under the change of variables, to the posterior resulting from a Jeffreys prior on the original set.

Recall that, in using (22) to derive (23), in which the prior on ρ is flat, the regressors in Z are treated as fixed. As Phillips (1991) explains, this conditioning is innocuous for the standard regression model but not for a time-series model, in which the expectation in (22) should reflect the stochastic nature of Z. For the two-equation model considered here, as in the AR(1) model, an exact Jeffreys prior depends on the sample size T and is complicated. As T grows large and $|\rho| < 1$, the limiting form of the Jeffreys prior is given by

$$p(b, \Sigma) \propto (1 - \rho^2)^{-1} \sigma_v^2 |\Sigma|^{-5/2}, \quad \rho \in (-1,\ 1), \tag{27}$$

as shown in the Appendix. For cases in which it is assumed that $|\rho| < 1$ and the exact likelihood in (25) is used to obtain posterior distributions, the limiting or "approximate" Jeffreys prior in (27) is entertained as an alternative to the flat prior in (26).

Whether or not a Jeffreys prior appropriately represents ignorance has long been a point of contention in Bayesian statistics, and this study has nothing to add in that regard. In any event, though, the prior in (27) assigns greater probability to values of ρ near 1 than does the flat prior on ρ in (26).[12] In applications where x_t is believed a priori to be highly

[11] Phillips explores the use of a Jeffreys prior for models in which stationarity is not imposed.

[12] Since ρ^2 appears in (27), greater prior probability is also assigned to values of ρ near -1. In the application considered here, modifying that prior with the restriction $0 \leq \rho < 1$ has essentially no effect no the results, since the values of the likelihood function are extremely small for ρ near -1.

autocorrelated, which is perhaps a reasonable belief for variables such as dividend yields and interest rates, the prior in (27) might be favored over one that is flat with respect to ρ. Leamer (1991), for example, discusses how aspects of such a non-flat prior can be appealing, even if one does not necessarily embrace the usual justifications for Jeffreys priors.

In the empirical analysis below, posterior distributions are computed for various combinations of the prior densities in (23), (26), and (27) and the likelihood functions in (20) and (25). The bounded flat prior in (26) can be combined with both the conditional likelihood in (20) and the exact likelihood in (25), whereas the unbounded flat prior in (23) is used only with the conditional likelihood, since the exact likelihood requires $|\rho| < 1$. The prior in (27) is combined only with the exact likelihood, since combining that prior with the conditional likelihood results in a non-integrable posterior density.[13]

3.2. Results

Table 2 reports moments of the posterior distributions for β obtained under the various combinations of priors and likelihoods described above. Also reported for each specification is the posterior probability that $\beta \leq 0$. Results are reported for the same four sample periods used in constructing Table 1. Details of the calculations are provided in the Appendix.

For the specification in Part A, which combines the prior in (23) with the conditional likelihood in (20), the posterior mean of β is equal to $\hat{\beta}$, and the posterior probability that $\beta \leq 0$ is virtually identical to the p-value in Table 1 computed in the standard regression setting. (There is a minor difference in the degrees of freedom.) In other words, even though the frequentist sampling moments and p-values computed under the standard assumptions depart substantially from the correct values, they nevertheless admit the standard Bayesian interpretation when the prior and likelihood are given by (23) and (20). Thus, for example, although the correct p-value for the hypothesis $\beta = 0$ is equal to 0.17 for the 1927-96 period (Table 1), the posterior probability that $\beta \leq 0$ is only 0.06. This observation is analogous to a similar point made by Sims (1988) for the AR(1) model.

The posterior probability that $\beta \leq 0$ can differ across the specifications in Parts A through

[13]For α, β, θ, and Σ set to any values (with Σ positive definite), let $\bar{L}$ denote the minimum value of the right-hand side of (20) for $\rho \in [-1, 1]$. (Since the likelihood, given the other parameters, is proportional to a normal density in ρ, $\bar{L}$ occurs at one of the endpoints.) Then the integral of the product of the right-hand sides of (27) and (20), with respect to $\rho \in (-1, 1)$, is bounded below by $\sigma_v^2 |\Sigma|^{-5/2} \bar{L} \int_{-1}^{1} (1 - \rho^2)^{-1} d\rho = \infty$. An integrable posterior density can (in principle) be obtained by instead using the conditional likelihood to obtain an exact Jeffreys prior, which depends on x_0 and T and is more complicated.

D of Table 2. For example, those probabilities range from 0.05 to 0.28 in Part D, whereas three of the four probabilities are 0.05 or less in Part C. In the 1977–96 period, the posterior probability that $\beta \leq 0$ is 0.26 in Part A but only 0.05 in Part C. Recall from Table 1, however, that the frequentist p-value for that period is 0.64. In general, although differences in the Bayesian posterior tail probabilities are clearly evident across the alternative specifications, none of those probabilities is nearly as large as the p-value for the same period.

The posterior means of β range between 0.19 and 0.23 for the overall 70-year period, but those differences seem modest, at least when compared to that period's bias in $\hat{\beta}$ (0.07). Larger differences emerge in the shorter periods. For example, in the 45-year period from 1952–96, the posterior mean in Part A exceeds that in Part D by 0.16 (0.44 versus 0.28), which is about the same as the bias in $\hat{\beta}$ for that period (0.18). In the 20-year period from 1977–96, although the differences across methods are not as large as the bias in $\hat{\beta}$ for that period (0.42), the posterior mean of β in Part C is twice the posterior mean in Part A (0.38 versus 0.19).

The posterior means of β obey a simple relation to the posterior means of ρ within a period. For all four Bayesian specifications,

$$\mathrm{E}\{\beta|\rho,\Sigma,D\} = \hat{\beta} + \frac{\sigma_{uv}}{\sigma_v^2}(\rho - \hat{\rho}), \tag{28}$$

as shown in the Appendix. Taking expectations of (28) with respect to ρ and Σ gives

$$\mathrm{E}\{\beta|D\} \approx \hat{\beta} + \mathrm{E}\left\{ \frac{\sigma_{uv}}{\sigma_v^2} \middle| D \right\} (\mathrm{E}\{\rho|D\} - \hat{\rho}). \tag{29}$$

The approximation error, which is equal to the posterior covariance between (σ_{uv}/σ_v^2) and ρ, is small for the samples used here, and the posterior mean of σ_{uv}/σ_v^2 is very similar across methods within a given sample period. For the regression of stock return on dividend yield, the posterior mean of σ_{uv}/σ_v^2 ranges roughly between -14 and -22 across the four sample periods. The negative relation in (28) produces a strong negative posterior correlation between β and ρ: that correlation ranges from -0.89 to -0.94 across the various methods and periods.

The relation in (29) links differences across methods in the posterior means of β to differences in the posterior means of ρ, and the latter differences can be traced to the alternative specifications of priors and likelihoods. For example, one regularity in Table 2 is that the posterior mean of β in Part C exceeds that in Part D in every period. Therefore, from (29), the posterior mean of ρ is lower for the specification in Part C than in Part D, and that ordering is consistent with the fact that the flat prior in Part C assigns less mass to regions

near $\rho = 1$ than does the approximate Jeffreys prior used in Part D. Another regularity suggested by (29) is that the posterior mean of β in Part B should be no less than that in Part A. Those specifications essentially differ only in that Part B rules out values of ρ above 1.0, so the posterior mean of ρ is lower than when such values are permitted in Part A. Given (29), the posterior mean of β should then be higher in Part B. In the first three sample periods, the differences between the posterior means in Parts A and B are neglibible, although consistent with the prediction. In the 1977–96 period, the posterior mean in Part B exceeds that in Part A by about 40% (0.27 versus 0.19).

The ordering of the posterior means in Parts B and C varies across sample periods. Those specifications share the same prior but have different likelihoods. The conditional likelihood used in Part B is multiplied by the density of the initial observation x_0, in (24), to obtain the exact likelihood in Part C. Including the density of x_0, which contains the parameters ρ, θ, and σ_v, affects the posterior mean of ρ, and thereby the posterior mean of β, in an unpredictable direction. As a result, the overall ordering of the posterior means of β differs across subperiods. For example, the posterior mean in Part A is greater than or equal to the other three posterior means in the 1952–96 period, but it is less than the other three in the 1977–96 period.

Figure 1 plots, for each sample period, the posterior mean of β versus the posterior mean of ρ based on the four specifications for the prior and likelihood used in Table 2. Also plotted are the MLE's of β and ρ based on the exact likelihood in (25) as well as bias-corrected OLS estimates. The latter are constructed by adjusting $\hat{\beta}$ for its bias, using Proposition 3, and then adjusting $\hat{\rho}$ for its bias, using Proposition 4. (As before, the values of ρ and Σ used in those calculations are set equal to the quantities obtained in the OLS estimation for each period.) Observe that, within a sample period, the six alternative estimates of β plot as a nearly perfect linear relation to the corresponding estimates of ρ. This result is predicted by (29) as well as two similar relations that govern the bias-corrected OLS estimates and the MLE's. The first of these follows directly from (17), which implies

$$\bar{\beta} = \hat{\beta} + \frac{\sigma_{uv}}{\sigma_v^2}\{\bar{\rho} - \hat{\rho}\}, \tag{30}$$

where "$\bar{}$" denotes a bias-corrected OLS estimator. The second relation, governing the MLE's based on the exact likelihood function L_e, is given by

$$\check{\beta} = \hat{\beta} + \frac{\check{\sigma}_{uv}}{\check{\sigma}_v^2}(\check{\rho} - \hat{\rho}), \tag{31}$$

where "$\check{}$" denotes an exact-likelihood MLE. (The Appendix contains a derivation.) The MLE for σ_{uv}/σ_v^2 in (31) is close to the posterior mean for that quantity in (29) as well as the

OLS-based estimate of that quantity used in applying (30). Thus, equations (29), (30), and (31) all represent essentially the same linear relation between estimates of β and ρ across methods. (Also note that the point $(\hat{\rho},\ \hat{\beta})$ obeys all three equations.)

As illustrated in Figure 1, differences across the methods in estimates of β can be ascribed to differences in estimates of ρ. The latter differences, often less than 0.01, might be viewed as negligible for many purposes, but when they are multiplied by σ_{uv}/σ_v^2, whose estimates range between -14 and -22 across the sample periods, the resulting differences in the estimates of β can be substantial. In 1952–96 period, for example, the posterior means of ρ lie between 0.980 and 0.988, while the posterior means of β range from 0.28 to 0.44. Similarly, in the 1977–96 period, the posterior means of ρ lie between 0.978 and 0.985, while the posterior means of β range from 0.19 to 0.38.

Observe also from Figure 1 that, in all four sample periods, the bias-corrected OLS estimate of β (point F) is less than any of the Bayesian posterior means (points A through D). The linear relation between estimates of β and ρ therefore implies that the bias-corrected estimate of ρ is greater than any of the posterior means of ρ. Even for the Bayesian specification in which $|\rho| < 1$ but much of the prior mass is assigned to values near unity (point D), the posterior mean of ρ is still less than the OLS estimate adjusted upward for its bias, which is also derived assuming $|\rho| < 1$. As noted in the previous section, the bias-corrected estimate of ρ in the last subperiod exceeds 1.0, while such an outcome is impossible for the posterior mean of ρ under any of the four Bayesian specifications entertained. The posterior mean under specification D is closest to the bias-corrected value in the 1952–96 sample period, which is the period used in the next section to analyze the asset-allocation decision.

The higher-order posterior moments in Table 2 reveal further characteristics of the wedge separating classical and Bayesian results in the current regression setting. Recall from Table 1 that the finite-sample distribution of $\hat{\beta}$ exhibits marked positive skewness and excess kurtosis in the regression of stock return on dividend yield. In contrast, although the posterior distribution of β has skewness in the 1952–96 and 1977–96 periods as high as 0.37 and 0.53 (part D), those values are still only one-third to one-half of the corresponding values in Table 1. Similarly, the kurtosis values for β in Table 2 all lie between 2.84 and 3.18, whereas the kurtosis values for $\hat{\beta}$ in Table 1 range from 3.84 to 5.83. In brief, the higher-order moments of the Bayesian posterior distributions in Parts B through D depart only modestly from the standard Bayesian-regression-model values (which are virtually identical to those in Part A), whereas the higher-order sampling moments of $\hat{\beta}$ depart substantially from their standard values.

4. Predictive distributions and asset allocation

4.1. Framework

A posterior distribution for the parameters in (1) and (3) implies a "predictive" distribution for future excess returns. Recall that y_T is the sample's most recent one-month excess return (continuously compounded), so the excess return over the K following periods is

$$y_{T+K,(K)} \equiv \sum_{k=1}^{K} y_{T+k}. \tag{32}$$

The predictive distribution of the K-period excess return is given by

$$p(y_{T+K,(K)}|D) = \int_{b,\Sigma} p(y_{T+K,(K)}|b,\Sigma,D)p(b,\Sigma|D)dbd\Sigma, \tag{33}$$

where $p(b,\Sigma|D)$ is the posterior density of b and Σ. In other words, $p(y_{T+K,(K)}|D)$ gives the probability distribution for the K-period excess return perceived by an investor at the end of period T. If the investor knew b and Σ, then the only relevant item from the sample would be x_T, the most recent observation of the predictive variable. When b and Σ are unknown, however, the investor uses all of the sample information to update his beliefs about those parameters, and the remaining parameter uncertainty, known as "estimation risk," is reflected in the predictive distribution of $y_{T+K,(K)}$.

In this section, the posterior distributions of b and Σ are explored in terms of their implied predictive distributions. An economic perspective on the predictive distributions is provided by exploring implications for asset allocation. For each posterior distribution, predictive distributions are obtained for hypothetical samples that have different values of x_T but produce the same posterior distribution for b and Σ as the actual sample. For each such hypothetical sample, the predictive distribution is unique. Varying the hypothetical samples in this manner and calculating the optimal asset allocation for each sample gives an economic characterization the sample evidence on return predictability. (The Appendix discusses details of the calculations involving the predictive distributions.)

Consider a hypothetical buy-and-hold investor who allocates invested wealth between stocks and cash (which earns a riskless interest rate). The investor faces one of the predictive distributions obtained here and is assumed to maximize the expected utility of wealth at the end of K periods. Utility is given by the iso-elastic function,

$$U(W_{T+K}) = \frac{1}{\gamma} W_{T+K}^{\gamma}, \tag{34}$$

with $\gamma \neq 0$. Kandel and Stambaugh (1996) suggest that the sensitivity of such an investor's stock allocation to a set of predictive variables provides an economically relevant metric by which to assess the strength of the empirical evidence on predictability. Kandel and Stambaugh confine their analysis to a single-period investment horizon, while Barberis (1999) extends their framework to analyze long-horizon asset allocations. The Bayesian econometric model used in both studies corresponds to the first of the specifications entertained here, in which the prior in (23) is combined with the conditional likelihood function in (20).[14] The asset allocations computed here for investment horizons of various lengths provide an economic perspective on the differences across the alternative Bayesian specifications.

For simplicity, the continuously compounded riskless return on cash in each future month is assumed to be known and equal to i_T, the current rate. The optimal stock allocation, ω, as a fraction of current wealth W_T, is the solution to

$$\max_{\omega} \mathrm{E}\left\{U(W_{T+K}) | D\right\}, \tag{35}$$

where

$$W_{T+K} = W_T \left[\omega \exp\{y_{T+K,(K)} + K i_T\} + (1-\omega)\exp\{K i_T\}\right]. \tag{36}$$

The expectation is taken with respect to the predictive distribution in (33). The stock allocation ω is confined to the interval (0, 1), i.e., short sales of stock or the riskless asset are precluded. The coefficient of relative risk aversion, $1-\gamma$, is set equal to 7. This value is chosen simply because it yields substantial allocations to stock while avoiding an excessive number of corner solutions at $\omega = 100\%$.

4.2. Results

Table 3 reports the optimal stock allocations implied by predictive distributions based on the 45-year period from 1952–96. The buy-and-hold investment horizons range from 1 month ($K = 1$) to 20 years ($K = 240$), and optimal stock allocations are computed for five different values of the most recent dividend yield, x_T, ranging from 1% to 6%. (The average dividend yield for the 1952–96 period is 3.8%.) Results are shown for three of the four specifications analyzed in Table 2. The results in Parts A and B of Table 2 are virtually identical for the 1952–96 period, so only the results using the specification in Part A are reported here. Also reported are optimal allocations for the case in which b and Σ are assumed to be known with certainty and set equal to the MLE's from the conditional likelihood (i.e., based on the

[14] Both of those studies include cases in which x_t is a vector of regressors, and that extension is discussed in the next section.

OLS estimation). This last case, provided for comparison, ignores estimation risk. Ignoring estimation risk has a substantial impact on the optimal stock allocation of a buy-and-hold investor at longer horizons, as observed previously by Barberis (1999). Note that, at a 20-year horizon, an investor with relative risk aversion equal to 7 who ignores estimation risk allocates 100% to stocks at all dividend yields, whereas an investor with the same risk aversion who incorporates estimation risk allocates at most 65% to stocks.

At horizons of one year or less, the stock allocation is increasing in the dividend yield for all methods that incorporate estimation risk, although there are substantial differences across methods. For example, at a current dividend yield of 5%, the one-year stock allocation is 100% in Part A but only 70% in Part C. The differences across methods can be nontrivial at the longer horizons as well. For example, when the current dividend yield is 5%, the stock allocation for a 10-year horizon is 76% in Part B but only 60% in Part C. At low values of the dividend yield, the stock allocation is generally increasing in the investment horizon, whereas that allocation is generally decreasing in the horizon at higher dividend yields. This effect is also noted by Barberis (1999). A result not previously reported is that, when estimation risk is incorporated, the optimal stock allocation is not monotonically increasing in the dividend yield at longer investment horizons. The various patterns in the optimal stock allocations can be understood to some degree by examining moments of the predictive distributions of $y_{T+K,(K)}$.

Tables 4 through 6 report the first three moments of the predictive distributions of $y_{T+K,(K)}$. The means and standard deviations in Tables 4 and 5 are expressed on an "annualized" basis. Specifically, the values in Table 4 are equal to $(12/K)$ times the mean of $y_{T+K,(K)}$, and the values reported in Table 5 are equal to $\sqrt{12/K}$ times the standard deviation of $y_{T+K,(K)}$. Observe that the expected returns in Table 4 are increasing in the most recent dividend yield, x_T. Because the degree of predictability of returns in more distant future months is less than in nearby months, the effect of the current dividend yield on future expected returns diminishes as the investment horizon grows. Even for 20-year returns, though, the differences between expected returns for $x_T = 3\%$ and $x_T = 5\%$ are typically 200 basis points per annum. That is, the persistence in dividend yield is sufficiently high so as to make the current dividend yield informative about expected returns well into the future. The patterns in the mean returns, by themselves, tend to make the optimal allocation increase in the dividend yield, with less sensitivity at longer horizons. As noted above, however, the optimal allocation need not increase in dividend yield at the longer horizons. A more complete explanation involves skewness, as will be discussed later.

The various methods that incorporate estimation risk produce different expected returns, although the differences are larger at the shorter investment horizons. At short horizons, the expected returns in Part C of Table 4 exhibit the least sensitivity to dividend yield, and those in Part A exhibit the greatest sensitivity. The lower sensitivity in Part C essentially reflects the lower posterior mean of β for that method in the 1952–96 period, as reported in Table 2. Even in that case, however, differences in the current dividend yield produce large differences in expected returns: increasing the dividend yield from 3% to 5% raises the expected 1-year return from 2.1% to 8.2%. Dividend yield affects even the 20-year expected return, as noted above, but the differences across methods are smaller than at shorter horizons. This closer agreement across methods at long horizons reflects in part the fact that future expected returns revert to their long-run unconditional mean, but it also reflects the negative relation between the posterior means of β and ρ discussed in the previous section. A lower value of β reduces the importance of x_T at all horizons, but a higher value of ρ increases the importance of x_T at longer horizons. Therefore, the negative association between the posterior means of β and ρ tends to mitigate the expected-return differences across methods at longer horizons.

The conditional likelihood function is used to obtain the predictive expected returns in Part A of Table 4, and that same conditional likelihood is used to obtain the MLE's used in constructing Part D. Comparing the results in Parts A and D reveals that estimation risk plays a negligible role in determining predictive expected returns. In contrast, a comparison of Parts A and D in Table 5, which reports predictive standard deviations, reveals nontrivial estimation risk, particularly at longer horizons. For example, when the current dividend yield x_T is 4%, the annualized standard deviation of the 20-year rate of return is more than 9% in Part A but only 6.5% in Part D. Moreover, in Part A, the effects of estimation risk are greater as the current dividend yield assumes extreme values. The latter effect reflects the fact that, although x_T enters the conditional expected return with greater importance when it is extreme, conditional on β and ρ, the uncertainty about those parameters also results in greater uncertainty about the conditional mean when x_T assumes more extreme values.

Comparing the standard deviations in Part A of Table 5 to those in Parts B and C reveals another effect of differences in prior beliefs about whether $|\rho| < 1$. In Parts B and C, where it is assumed that $|\rho| < 1$, the predictive distribution of future returns is stationary, and the annualized standard deviation decreases with the investment horizon. In Part A, the predictive distribution of future returns is nonstationary, because the posterior density of ρ assigns positive mass to $|\rho| > 1$. Moreover, for the sample analyzed, sufficient posterior mass is assigned to $\rho > 1$ so as to make the effects of nonstationarity evident at the 20-year horizon. Observe in Part A that, for some values of x_T, the standard deviations for the

20-year horizon are higher than for the 10-year horizon, in contrast to the results in Parts B and C.

Recall that when estimation risk is incorporated, the stock allocation is often decreasing in dividend yield at the 20-year horizon, even though the expected 20-year return is monotonically increasing in dividend yield. The standard deviations in Table 5 do not appear to resolve this seeming contradiction, since the 20-year standard deviations are U-shaped with respect to dividend yield. That is, in all three methods that incorporate estimation risk, the stock allocation at a dividend yield of 3% is higher than the allocation at a yield of 5%, even though the latter value is associated with a higher mean and, in Parts B and C, a lower standard deviation of $y_{T+K,(K)}$.

4.3. Skewness and the role of uncertainty about ρ

A clue to the patterns in the long-horizon stock allocations is provided by the skewness coefficients in Table 6. Observe that, at the longer horizons, the predictive skewness of $y_{T+K,(K)}$ is positive at low dividend yields and negative at high yields. In Part C, for example, the 20-year return has skewness equal to 0.6 at a 2% dividend yield and -0.8 at a 6% dividend yield. A similar pattern occurs in Parts A and B, except that the magnitudes are much larger in Part A, where values of $|\rho|$ greater than 1 are permitted.

Positive skewness in $y_{T+K,(K)}$ can lead to a higher stock allocation than obtained with negative skewness, holding other moments constant. If r denotes the continuously compounded return on the investor's overall portfolio, so $W_{T+K} = W_T \exp(r)$, then a third-order approximation for expected utility is given by

$$\mathrm{E}\{U(W_{T+K})\} = \frac{W_T^\gamma}{\gamma} \exp[\gamma \bar{r}] \left[1 + \frac{\gamma}{2}\mathrm{var}\{r\} + \frac{\gamma^2}{6}\mathrm{E}(r-\bar{r})^3 + \mathrm{E}\{O[(r-\bar{r})^4]\}\right], \qquad (37)$$

where $\bar{r} \equiv \mathrm{E}\{r\}$. Thus, expected utility is increasing in the skewness of r. For a given stock allocation ω, the skewness in $y_{T+K,(K)}$ does not necessarily translate to skewness in r. In the current problem, it appears from numerical investigation that, for a given value of x_T, the skewness of r at long horizons is decreasing in ω. With low values of x_T, for which $y_{T+K,(K)}$ has positive skewness, the skewness of r is positive at all levels of ω but largest at the smallest ω values. With high values of x_T, for which $y_{T+K,(K)}$ is negatively skewed, the skewness of r is also positive for small values of ω but then becomes negative as ω increases. In general, the pattern in the skewness of $y_{T+K,(K)}$ in Table 6 tends to work in opposition to the pattern in the expected return, and the result is an optimal stock allocation that

can actually be higher at lower values of the current dividend yield, i.e., at lower expected returns.

The skewness in $y_{T+K,(K)}$ can be traced to estimation risk. For a given value of x_T, a draw from the predictive distribution of $y_{T+K,(K)}$ can be written as

$$y_{T+K,(K)} = c + d(x_T - \bar{x}) + s\eta, \tag{38}$$

where $\bar{x}$ denotes the sample mean of x_t $(= (1/T)\iota_T' x_{(\ell)})$, and η is a standard normal $(0, 1)$ variate that is independent of b and Σ. The coefficients c, d, and s are functions of the unknown parameters b and Σ, which are drawn from the posterior distribution $p(b, \Sigma|D)$, and c also depends on the known sample quantity $\bar{x}$. (Expressions for c, d, and s are provided in the Appendix.) Denote the conditional mean of $y_{T+K,(K)}$ given b, Σ, and x_T as

$$e = c + d(x_T - \bar{x}), \tag{39}$$

and define that quantity's deviation from its posterior mean as

$$\begin{aligned} \tilde{e} &= [c - \mathrm{E}\{c|D\}] + [d - \mathrm{E}\{d|D\}](x_T - \bar{x}) \\ &= e - \mathrm{E}\{e|D\}. \end{aligned} \tag{40}$$

The predictive third moment of $y_{T+K,(K)}$ can then be written as

$$\mathrm{E}\{(y_{T+K,(K)} - \mathrm{E}\{y_{T+K,(K)}|D\})^3|D\} = \mathrm{E}\{\tilde{e}^3|D\} + 3\mathrm{E}\{\tilde{e}s^2|D\}, \tag{41}$$

relying on the properties of η stated above. Since each skewness value reported in Table 6 is simply the third moment in (41) divided by the predictive variance to the power 3/2, the sign of the skewness is the same as that of (41).

Uncertainty about ρ plays a key role in explaining the skewness patterns. Consider the specification in Part A of Table 6, where skewness and its effects on asset allocation (in Table 3) are most pronounced. Figure 2 displays the marginal posterior density of ρ (upper left graph) as well as graphs that plot draws of ρ versus draws of the various quantities in equation (38) for $K = 240$ (20 years). As before, the quantities are annualized, so that c, d, and e are multiplied by $(12/K)$ and s is multiplied by $\sqrt{12/K}$ (but the scales are decimal values, not percents). For relatively high draws of ρ, especially those greater than 1, observe that d takes large negative values (middle left graph). As a result, for high draws of ρ, e takes large positive values for low values of x_T and large negative values for high values of x_T (by equation (39)). These two scenarios are illustrated in Figure 2 for $x_T = 2\%$ (bottom left graph) and $x_T = 6\%$ (bottom right graph). In other words, $\mathrm{E}\{\tilde{e}^3|D\}$, the first term on the

right-hand side of (41), is positive for low x_T's and negative for high x_T's, and this pattern is the same as that observed for the predictive skewness in Table 6. The use of $\bar{x}$ (about 3.8%) as the reference value in (38) is somewhat arbitrary, but with this simple choice the intercept c exhibits only minor skewness for large values of ρ, thereby allowing the slope coefficient d to isolate the main effect of uncertainty about ρ.

The posterior uncertainty about ρ can be sufficient to assign small but nontrivial probability to high values of ρ, even values above 1 in the specification used to construct Figure 2. A higher value of ρ implies that x_T has a more persistent effect on future mean returns, so the absolute value of d is then larger, given β. To understand why the extreme d values are typically negative, as illustrated in Figure 2, recall that the posterior correlation between ρ and β is strongly negative, equal to -0.94 in this example. Hence, if ρ is high, β is likely to be low, negative in fact, so the extreme values of d tend to be negative. Of course, d has a positive posterior mean, which is computed by averaging over all posterior draws of ρ and β. Thus, when x_T is low, the predictive mean of $y_{T+K,(K)}$ is also low, as demonstrated in Table 4. If there is a chance, however, that the value of ρ is higher than, say, its posterior mean, there is also a chance that the true mean e of the long-horizon return is substantially higher than its (low) posterior mean, so e is positively skewed. Similarly, when x_T is high, there is a chance that e is substantially lower than its (high) posterior mean, so e is negatively skewed.

Also observe in Figure 2 (middle right graph) that high values of ρ produce large values of s in (38), where s is the standard deviation of $y_{T+K,(K)}$ conditional on b, Σ, and x_T. When x_T is low, high values of ρ produce high volatility accompanied by large positive values of the conditional mean, thereby adding to the positive skewness in the predictive distribution. In other words, a low x_T produces a positive value for $\mathrm{E}\{\tilde{e}s^2|D\}$, which is proportional to the second term on the right-hand side of (41). Similarly, a high x_t produces a negative value for $\mathrm{E}\{\tilde{e}s^2|D\}$. Thus, the positive association between ρ and the conditional volatility amplifies the skewness effect produced by the behavior of the conditional mean.

The explanation for the skewness patterns in Parts B and C of Table 6 follows the same lines as detailed above for Part A. Precisely the same reasoning applies, except that ρ cannot exceed 1 in Parts B and C. The effects are hence weaker but nevertheless present. (Note that truncating the graphs in Figure 2 at $\rho = 1$ still leaves some of the patterns.) In general, uncertainty about ρ produces positive skewness for low values of dividend yield and negative skewness for high values.

5. Extensions to multiple predictive variables

The predictive regressions considered in the preceding sections contain a single independent variable, but much of the analysis can be generalized to settings in which x_t in (1) is a vector instead of a scalar. A tractable model for such a generalization assumes the $N \times 1$ vector h_t follows a first-order vector autoregression (VAR),

$$h_t = \phi_0 + \Phi h_{t-1} + e_t, \tag{42}$$

where e_t is an independent realization from a multivariate normal distribution with mean zero and covariance matrix Σ (now $N \times N$). With multiple predictive variables, the excess return y_t is simply the first element of h_t, so the vector of predictive variables, in general, can contain the lagged value of y_t. The first row of Φ contains the slope coefficients in the regression of y_t on the N predictive variables. Note that $\mathrm{E}\{e_t|h_{t-1}, h_{t-2}, \ldots\} = 0$ but $\mathrm{E}\{e_t|h_s, h_w\} \neq 0$ for $s < t \leq w$, and the latter condition corresponds to (2). The two-equation model comprising (1) and (3) can be represented as a special case of (42) in which $h_t = (y_t \; x_t)'$, $e_t = (u_t \; v_t)'$, $\phi_0 = (\alpha \; \theta)'$, and Φ has zeros in the first column and (1, 2) and (2, 2) elements equal to β and ρ.

The first Bayesian specification, in which the prior in (23) is combined with the conditional likelihood in (20), extends immediately to the above VAR with the quantities appropriately redefined. Specifically, let

$$z = \mathrm{vec}\,([h_1 \; h_2 \; \cdots \; h_T]), \tag{43}$$

$$X = [\iota_T \;\; (h_0 \; h_1 \; \cdots \; h_{T-1})'], \tag{44}$$

$$b = \mathrm{vec}\,([\phi_0 \; \Phi]'), \tag{45}$$

and

$$Z = (I_N \otimes X), \tag{46}$$

where vec( ) forms a column vector by stacking successive columns of the matrix. The right-hand side of (20) is then the conditional likelihood for the VAR in (42), under the assumption that the vector of initial observations h_0 is nonstochastic.[15] When modified for the case of N equations, the prior in (23) becomes

$$p(b, \Sigma) \propto |\Sigma|^{-(N+1)/2}. \tag{47}$$

[15]As explained by Hamilton (1994, p.358), for example, a model with lagged dependent variables (such as the VAR) can be analyzed as a standard Bayesian multivariate regression model if the "pre-sample" observations are assumed to be deterministic.

This Bayesian VAR specification with multiple predictive variables is used in the analyses of asset allocation by Kandel and Stambaugh (1996) and Barberis (1999) and in an analysis of currency hedging by Bauer (1998). In this specification, the prior for each element of Φ is flat over the real line, as is the prior for ρ under the corresponding specification in the single-variable setting. That is, the prior in (47) does not impose covariance-stationarity, since that condition requires the eigenvalues of Φ to lie inside the unit circle (e.g., Hamilton, 1994, p. 259). The latter condition, represented here by the notation $\|\Phi\| < 1$, is equivalent to requiring $|\rho| < 1$ with one predictive variable.

Recall that in the single-variable setting in Section 3, two alternative Bayesian specifications are considered, each of which imposes covariance-stationarity and uses the exact likelihood in (25). In the N-variable setting, the exact likelihood is defined by the assumption that h_0 is drawn from its unconditional distribution. From (42), that distribution has mean

$$\mu_h = (I_N - \Phi)^{-1}\phi_0 \tag{48}$$

and variance-covariance matrix V_h satisfying $V_h = \Phi V_h \Phi' + \Sigma$, which can be solved in terms of b and Σ (Hamilton, 1994, p. 265) to yield

$$\text{vec}\,(V_h) = [I_{N^2} - (\Phi \otimes \Phi)]^{-1}\text{vec}\,(\Sigma). \tag{49}$$

The exact likelihood in the N-variable case is

$$L_e(b, \Sigma; D) = p(z, h_0|b, \Sigma) = p(z|h_0, b, \Sigma)p(h_0|b, \Sigma), \tag{50}$$

where $p(z|h_0, b, \Sigma)$ is given by the right-hand side of (20) and

$$p(h_0|b, \Sigma) = (2\pi)^{-1/2}|V_h|^{-1/2}\exp\left\{-\frac{1}{2}(h_0 - \mu_h)'V_h^{-1}(h_0 - \mu_h)\right\}. \tag{51}$$

The prior in (26), which keeps a flat prior on ρ but simply imposes the stationarity restriction on the prior on (23), can be similarly adapted here. That is, the prior in (47) can be applied to the regions of the parameter space in which $\|\Phi\| < 1$, so that the prior density is zero elsewhere. The approximate Jeffreys prior in (27), when generalized to the N-variable setting, becomes

$$p(b, \Sigma) \propto |V_h|^{N/2}|\Sigma|^{-(N+1)}, \tag{52}$$

as shown in the Appendix. The techniques described in the Appendix for obtaining the posterior and predictive distributions in the single-variable case extend in a straightforward manner to the N-variable case.

Extending the analytical results for the finite-sample properties of the OLS estimator is less straightforward, since Propositions 1 through 4 do not appear to generalize easily to N lagged stochastic variables. The problem in the N-variable case can be characterized as analyzing the finite-sample distribution of the OLS estimator

$$[\ \hat{\phi}_0 \quad \hat{\Phi}\ ]' = (X'X)^{-1}X'H, \tag{53}$$

where $H = [h_1\ h_2\ \cdots h_T]'$. Note that the first row of $\hat{\Phi}$ contains the OLS estimates of the slope coefficients in a multiple regression of the return (y_t) on the N lagged predictive variables. When $\|\Phi\| < 1$ and e_t obeys the normal distribution as above, Nicholls and Pope (1988) show that an approximation to the bias in $\hat{\Phi}$ is given by

$$\mathrm{E}\{\hat{\Phi}\} - \Phi = \Sigma \left[(I_N - \Phi)^{-1} + \Phi(I_N - \Phi^2)^{-1} + \sum_{\lambda \in s(\Phi)} \lambda (I_N - \lambda\Phi)^{-1} \right] V_h^{-1} + O(T^{-3/2}), \tag{54}$$

where the notation $\sum_{\lambda \in s(\Phi)}$ denotes summation over the eigenvalues of Φ, with each term repeated as many times as the multiplicity of the eigenvalue λ.

6. Conclusions

When the innovation in a lagged stochastic regressor is correlated with the regression disturbance, the OLS estimator can exhibit finite-sample properties that deviate sharply from those in the standard regression setting. One example of such a regression occurs when the aggregate stock portfolio's excess rate of return is regressed on its lagged dividend yield. In that application, the bias in the OLS slope coefficient ranges from one-third of the OLS estimate in the 1927–96 period to more than three times the OLS estimate in the 1977–96 period. The finite-sample p-values for a one-tailed test of the zero-slope hypothesis range between 0.17 and 0.64 across the various periods considered, and those p-values are substantially larger than the p-values computed incorrectly using the standard regression model.

In the results obtained here for the dividend-yield regression, the p-value for the zero-slope hypothesis exceeds the Bayesian posterior probability that the regression slope is less than or equal to zero. In the 1952–96 period, for example, the p-value equals 0.15, so a classical test would accept the zero-slope hypothesis at conventional significance levels. In contrast, the posterior probability that the slope is less than or equal to zero ranges between 0.01 and 0.05, depending on the specification of the likelihood and prior. The potential conflict between frequentist and Bayesian inference assumes greater prominence with a lagged stochastic regressor, since the p-value and the posterior tail probability coincide in the standard regression setting.

Bayesian posterior distributions for the parameters of the regression model exhibit sensitivity to whether (i) the initial observation of the regressor is viewed as fixed or stochastic, (ii) the regressor is assumed to be stationary, and (iii) a "flat" prior or a Jeffreys prior is employed. The OLS estimator of the regression coefficient vector is also the posterior mean when the initial observation is fixed and the prior for the autoregressive coefficient of the regressor is flat over the real line (allowing nonstationarity). One alternative specification employs a Jeffreys prior and assumes that the initial observation is a stochastic realization from a stationary process for the regressor. The Jeffreys prior, also intended to be noninformative, assigns higher posterior density to autoregressive coefficients near unity. In the 1952–96 period, the posterior mean of the regression slope is more than 50% higher with the first specification than with the second. Such sensitivity underscores the finite-sample nature of the regression problem considered here. Moreover, this sensitivity is not limited to a Bayesian setting. In the same 1952–96 period, for example, the OLS slope estimate is 27% higher than the maximum-likelihood estimate computed under the assumption that the initial observation of the regressor is a stochastic realization from a stationary process.

The regression of excess stock returns on dividend yield is used as an illustration here in part because the posterior distributions for the parameters can be used to compute predictive distributions for future excess stock returns. The predictive distribution, which incorporates "estimation risk" arising from parameter uncertainty, can then be used to compute the optimal portfolio for a buy-and-hold investor facing a stocks-versus-cash allocation decision. These computations provide an economic setting for comparing the various econometric specifications, and the differences across specifications can be economically important. In an example using the 1952–96 period, if the most recent dividend yield is 5%, an investor with a 5-year horizon and relative risk aversion equal to 7 chooses a stock allocation between 68% and 86%, depending on the specifications of the prior and the likelihood.

The asset-allocation results also reveal a new insight into the potential role of estimation risk in long-horizon investing. In particular, at longer investment horizons, the optimal buy-and-hold stock allocation can be higher at low values of the current dividend yield than at high values, even though the long-horizon stock return has a lower mean at the low dividend yield and can have at least as high a variance. This result can be traced to skewness in long-horizon stock returns arising from uncertainty about parameters, particularly the autoregressive coefficient of dividend yield. The skewness in the predictive distribution of returns is positive at low dividend yields and negative at high yields, and the effect of this skewness can be strong enough to produce a negative association between the optimal stock allocation and dividend yield.

Appendix

A.1. Proof of Proposition 1

Define $\bar{x} = (1/T)\iota_T' x_{(\ell)}$, and observe that

$$\begin{aligned}
\hat{\beta} &= \frac{(x_{(\ell)} - \iota_T \bar{x})' y}{(x_{(\ell)} - \iota_T \bar{x})'(x_{(\ell)} - \iota_T \bar{x})} \\
&= \frac{x_{(\ell)}' M y}{x_{(\ell)}' M x_{(\ell)}} \\
&= \beta + \frac{x_{(\ell)}' M u}{x_{(\ell)}' M x_{(\ell)}} \\
&= \beta + \frac{(x_{(\ell)} - \mu_x \iota_T)' M u}{(x_{(\ell)} - \mu_x \iota_T)' M (x_{(\ell)} - \mu_x \iota_T)} \\
&= \beta + \frac{w' A w}{w' B w}. \qquad \text{(A.1)}
\end{aligned}$$

The second equation uses the property $M^2 = M$, and the fourth equation uses the property $\iota_T' M = 0$. Clearly $\mathrm{E}\{w\} = 0$, and it is straightforward to verify that $\mathrm{cov}\{w,\ w'\} = \Omega$, as defined in the proposition. Note that α, β, and θ do not affect the distribution of $\hat{\beta} - \beta$, since those parameters do not enter Ω, A, or B.

A.2. Proof of Proposition 2

From (3) and the definition of w in Proposition 1, normality of $(u_t\ v_t)'$ for all t implies normality of w. Observe, using (6), that

$$\begin{aligned}
\mathrm{Prob}\{\hat{\beta} > \beta_0\} &= \mathrm{Prob}\{\beta + \frac{w' A w}{w' B w} > \beta_0\} \\
&= \mathrm{Prob}\{w' A w > (\beta_0 - \beta) w' B w\} \\
&= \mathrm{Prob}\{w' C w > 0\}, \qquad \text{(A.2)}
\end{aligned}$$

where $C = A - (\beta_0 - \beta)B$. Imhof (1961) gives a method, based on inversion of the characteristic function, for computing $\mathrm{Prob}\{w'Cw > c\}$, where w obeys a multivariate normal distribution, possibly with nonzero mean, and C is an indefinite matrix. The result in (9) is a direct application of Imhof's equation (3.2).

A.3. Proof of Proposition 3

Magnus (1986, Theorem 6) derives $\mathrm{E}\{[w'Aw/w'Bw]^s\}$, where A is a symmetric matrix, B is a positive semidefinite matrix of rank $r \geq 1$, and the $n \times 1$ vector w obeys a normal

distribution with mean μ and positive definite covariance matrix $\Omega = LL'$. His theorem is as follows. Let P be an orthogonal $n \times n$ matrix and Λ a diagonal $n \times n$ matrix such that $P'L'BLP = \Lambda$ and $P'P = I_n$. Then, provided the expectation exists (see below), for $s = 1, 2, \ldots,$

$$\begin{aligned} \mathrm{E}\left\{\left[\frac{w'Aw}{w'Bw}\right]^s\right\} &= s2^s \exp\{-\frac{1}{2}\mu'\Omega\mu\} \sum_i \gamma_s(\nu_i) \\ &\quad \times \int_0^\infty q^{s-1}|\Delta| \exp\{\frac{1}{2}\xi\xi'\} \prod_{j=1}^{s} (\mathrm{tr}\, R^j + j\xi' R^j \xi)^{n_{ij}} dq, \end{aligned} \tag{A.3}$$

where $\Delta = (I_n + 2q\Lambda)^{-1/2}$, $R = \Delta P'L'ALP$, $\xi = \Delta P'L^{-1}\mu$, and the summation is over all vectors $\nu_i = (n_{i1}, n_{i2}, \ldots, n_{is})$ whose s elements are non-negative integers satisfying $\sum_{j=1}^{s} j n_{ij} = s$, with

$$\gamma_s(\nu_i) = \prod_{j=1}^{s} [n_{ij}!(2j)^{n_{ij}}]^{-1}. \tag{A.4}$$

The result in (10) then follows directly from Proposition 1, with $n = 2T$ and $\mu = 0$.

If $r \leq n - 1$ and Q is an $n \times (n - r)$ matrix of full column rank $n - r$ such that

$$L'BLQ = 0, \tag{A.5}$$

then $\mathrm{E}\{[w'Aw/w'Bw]^s\}$ exists for $0 \leq s < r$ under the condition $Q'L'ALQ = 0$ (Magnus, 1986, Theorem 7).[16] In this application, the rank of B equals $T - 1$ (the rank of M), so Q is a $2T \times (T + 1)$ matrix. Let $L' = [L_1' \; L_2']$, where L_1 and L_2 are both $T \times 2T$ matrices. From (8), $L'BL = L_2'ML_2$ so (A.5) implies $L_2'ML_2Q = 0$, and since L_2 has full row rank,

$$ML_2Q = 0. \tag{A.6}$$

From (8),

$$L'AL = (1/2)(L_1'ML_2 + L_2'ML_1) \tag{A.7}$$

so (A.6) implies $Q'L'ALQ = 0$.

A.4. Proof of Proposition 4

Let $b_1 = (\alpha \; \beta)'$, $b_2 = (\theta \; \rho)'$, $\hat{b}_1 = (\hat{\alpha} \; \hat{\beta})'$, and $\hat{b}_2 = (\hat{\theta} \; \hat{\rho})'$. Equations (5) and (16) imply

$$\hat{b}_1 - b_1 = (X'X)^{-1}X'u \tag{A.8}$$

and

$$\hat{b}_2 - b_2 = (X'X)^{-1}X'v, \tag{A.9}$$

[16] Magnus's theorem contains an alternative condition for moments to exist for $s \geq r$ as well, but that condition is not satisfied here.

where u is defined previously and $v = (v_1 \ \ldots \ v_T)'$. Decompose u as

$$u = \frac{\sigma_{uv}}{\sigma_v^2} v + \epsilon, \tag{A.10}$$

with $\mathrm{E}\{\epsilon|v\} = 0$ implied by the i.i.d. normality assumption, so

$$\mathrm{E}\{\epsilon|X\} = \mathrm{E}\{\epsilon|x_0, v_1, \ldots, v_{T-1}\} = 0. \tag{A.11}$$

Substituting from (A.10) into (A.8) gives

$$\begin{aligned} \hat{b}_1 - b_1 &= \frac{\sigma_{uv}}{\sigma_v^2}(X'X)^{-1}X'v + (X'X)^{-1}X'\epsilon \\ &= \frac{\sigma_{uv}}{\sigma_v^2}(\hat{b}_2 - b_2) + (X'X)^{-1}X'\epsilon, \end{aligned} \tag{A.12}$$

where the second equality uses (A.9). Taking expectations, using (A.11), gives

$$\mathrm{E}\{\hat{b}_1 - b_1\} = \frac{\sigma_{uv}}{\sigma_v^2}\mathrm{E}\{\hat{b}_2 - b_2\}, \tag{A.13}$$

and (17) is the second row of the vector equation in (A.13).

A.5. Derivation of the Jeffreys prior in (27)

For the stationary AR(1) model, Zellner (1971, pp. 216–220) obtains an "approximate" Jeffreys prior by retaining only the terms that are of the highest order of T when applying (22) to the exact likelihood.[17] Such an approach is equivalent to computing the Jeffreys prior for the conditional likelihood and taking the expectation in (22) with the initial observation x_0 assumed to be stochastic and drawn for its unconditional distribution. The same equivalence occurs for the two-equation model analyzed here. In implementing the latter approach, it is convenient to derive the joint prior $p(b, \Sigma^{-1})$ and then make the transformation from Σ^{-1} to Σ. The log-likelihood for (20) is given by

$$\ell \equiv \log L_c(b, \Sigma; z, x_0) = -\frac{T}{2}\log|\Sigma| - \frac{1}{2}(z - Zb)'(\Sigma^{-1} \otimes I_T)(z - Zb). \tag{A.14}$$

Let $\zeta \equiv (\sigma^{11} \ \sigma^{12} \ \sigma^{22})'$, where σ^{ij} denotes the (i,j) element of Σ^{-1}. Following (22),

$$p(b, \Sigma^{-1}) \propto \left| -\mathrm{E}\left\{ \begin{array}{cc} \frac{\partial^2 \ell}{\partial b \partial b'} & \frac{\partial^2 \ell}{\partial b \partial \zeta'} \\ \frac{\partial^2 \ell}{\partial \zeta \partial b'} & \frac{\partial^2 \ell}{\partial \zeta \partial \zeta'} \end{array} \right\} \right|^{1/2}. \tag{A.15}$$

Observe

$$\begin{aligned} \frac{\partial \ell}{\partial b} &= \frac{1}{2}Z'(\Sigma^{-1} \otimes I_T)(z - Zb) \\ &= \frac{1}{2}Z'(\Sigma^{-1} \otimes I_T)\begin{bmatrix} u \\ v \end{bmatrix}, \end{aligned} \tag{A.16}$$

[17] See Uhlig (1994) for exact Jeffreys priors for the AR(1) model.

$$\begin{aligned}\frac{\partial^2 \ell}{\partial b \partial b'} &= -\frac{1}{2} Z'(\Sigma^{-1} \otimes I_T) Z \\ &= -\frac{1}{2}\left(\Sigma^{-1} \otimes X'X\right) \end{aligned} \tag{A.17}$$

and

$$\frac{\partial^2 \ell}{\partial b \partial \zeta'} = \frac{1}{2}(I_2 \otimes X') \begin{bmatrix} u & v & 0 \\ 0 & u & v \end{bmatrix}. \tag{A.18}$$

Taking the expectations of (A.17) and (A.18) with respect to $p(z, x_0|b, \Sigma)$ gives

$$\mathrm{E}\left\{\frac{\partial^2 \ell}{\partial b \partial b'}\right\} = -\frac{T}{2}\left(\Sigma^{-1} \otimes \Psi\right), \tag{A.19}$$

where

$$\Psi = \frac{1}{T}\mathrm{E}\{X'X\} = \begin{bmatrix} 1 & \theta/(1-\rho) \\ \theta/(1-\rho) & \sigma_v^2/(1-\rho^2) + \theta^2/(1-\rho)^2 \end{bmatrix}, \tag{A.20}$$

and

$$\mathrm{E}\left\{\frac{\partial^2 \ell}{\partial b \partial \zeta'}\right\} = 0. \tag{A.21}$$

Also observe

$$\mathrm{E}\left\{\frac{\partial^2 \ell}{\partial \zeta \partial \zeta'}\right\} = \frac{\partial^2 \ell}{\partial \zeta \partial \zeta'} = -\frac{T}{2}\frac{\partial^2 \log|\Sigma|}{\partial \zeta \partial \zeta'}. \tag{A.22}$$

Substituting from (A.19), (A.21), and (A.22) into (A.15) gives

$$\begin{aligned} p(b, \Sigma^{-1}) &\propto \left| \begin{matrix} \Sigma^{-1} \otimes \Psi & 0 \\ 0 & \frac{\partial^2 \log|\Sigma|}{\partial \zeta \partial \zeta'} \end{matrix} \right|^{1/2} \\ &\propto \left|\Sigma^{-1} \otimes \Psi\right|^{1/2} \left|\frac{\partial^2 \log|\Sigma|}{\partial \zeta \partial \zeta'}\right|^{1/2} \\ &= \left(|\Sigma|^{-2}|\Psi|^2\right)^{1/2}\left(|\Sigma|^3\right)^{1/2} \\ &= |\Psi||\Sigma|^{1/2} \\ &= (1-\rho^2)^{-1}\sigma_v^2|\Sigma|^{1/2}. \end{aligned} \tag{A.23}$$

The Jacobian of the transformation from Σ^{-1} to Σ is $|\Sigma|^{-3}$ (see Box and Tiao, 1973, p. 474), and multiplying (A.23) by that quantity gives (27).[18]

In a standard multivariate regression setting, a common practice is to apply (22) separately for b and Σ (e.g., Zellner, 1971, chapter 8), following the suggestion by Jeffreys (1961) to treat location and scale parameters separately in multiparameter settings. As Jeffreys notes, treating location and scale parameters jointly can result in unappealing degrees-of-freedom properties, such as the observation that, in the simplest i.i.d. normal univariate

[18]The Jacobian of the transformation from Σ^{-1} to Σ, as well as the determinant of the derivative matrix for $\log|\Sigma|$ in (A.23), follow from results in Box and Tiao (1973, pp. 474–475).

setting, the degrees of freedom in the posterior for the variance is unaffected by whether the mean is known or unknown (see, for example, Bernardo and Smith, 1994, chapter 5). In a time-series setting, the dichotomy between location and scale parameters is blurred. For example, ρ affects the conditional mean as well as the unconditional variance of x_t. Phillips (1991) argues that, in the AR(1) model, the usual degrees-of-freedom criticism does not apply to the multiparameter Jeffreys prior. In the current setting, applying (22) separately for b and Σ results in the prior

$$p(b, \Sigma) \propto (1 - \rho^2)^{-1} |\Sigma|^{-3/2}, \tag{A.24}$$

and the use of this alternative prior produces only negligible changes from the results obtained using (27), which has the same marginal prior on ρ.

A.6. Calculation of the posterior distributions

With the likelihood in (20) and the prior in (23), the posterior distribution for b and Σ follows from standard results for the multivariate regression model (e.g., Zellner, 1971, chapter 8). Specifically, Σ^{-1} obeys a Wishart distribution with $T-2$ degrees of freedom and parameter matrix S, where $S = (Y - X\hat{B})'(Y - X\hat{B})$, $Y = [y\ x]$, B is a 2×2 matrix with first row $(\alpha\ \theta)$ and second row $(\beta\ \rho)$, and $\hat{B}$ is the same reshaping of $\hat{b}$. The conditional distribution of b given Σ is normal with mean $\hat{b}$ and covariance matrix $\Sigma \otimes (X'X)^{-1}$. Those distributions are used to generate 100,000 independent draws of b and Σ, which are in turn used in generating draws from the predictive distribution of multiperiod returns (explained below). The marginal posterior distribution for β is Student t with $T-3$ degrees of freedom,

$$\mathrm{E}\{\beta|D\} = \hat{\beta}, \tag{A.25}$$

$$\mathrm{var}\{\beta|D\} = \frac{1}{T-5}\frac{\hat{\sigma}_u^2}{\hat{\sigma}_x^2}, \tag{A.26}$$

skewness equal to zero, and kurtosis equal to $3[1+2/(T-7)]$, where $\hat{\sigma}_x^2 = (1/T)\sum_{t=1}^{T}(x_{t-1} - \bar{x})^2$ and $\hat{\sigma}_u^2 = (1/T)\sum_{t=1}^{T}(y_t - \hat{\alpha} - \hat{\beta}x_{t-1})^2$. The values in Part A of Table 2 are based on the latter results.

With the likelihood in (20) and the prior in (26), the joint posterior density for b and Σ is proportional to the joint density described in the first case above multiplied by an indicator function equal to 1.0 if $|\rho| < 1$ and zero otherwise. Draws of $(\Sigma^{-1},\ b)$ are generated from the Wishart and conditional normal distributions described above and then retained only if $|\rho| < 1$. The values in Part B of Table 2 are based on 100,000 retained draws.

With the likelihood in (25) and the prior in (26), the joint posterior density for b and Σ is given by

$$p(b,\Sigma|D) \propto |\Sigma|^{-(T+5)/2}\exp\left\{-\frac{1}{2}(z-Zb)'(\Sigma^{-1}\otimes I_T)(z-Zb)\right\} \times\left(\frac{1-\rho^2}{\sigma_v^2}\right)^{1/2}\exp\left\{-\frac{1-\rho^2}{2\sigma_v^2}\left(x_0-\frac{\theta}{1-\rho}\right)^2\right\},\ \rho\in(-1,\ 1). \quad \text{(A.27)}$$

Integrating (A.27) analytically to obtain the marginal posterior density $p(\beta|D)$ does not appear to be feasible. Instead, that posterior density is obtained using the Metropolis-Hastings (MH) algorithm, a Markov chain Monte Carlo procedure introduced by Metropolis et al. (1953) and generalized by Hastings (1970).[19] A sequence of values for $(b\ \Sigma)$ is constructed by making "candidate" draws from a "proposal" density and then accepting a new candidate or retaining the previous value based on the MH rule that assures the resulting sequence for $(b,\ \Sigma)$ forms a Markov chain whose invariant distribution is the "target" posterior density in (27).

The MH algorithm is implemented with b and Σ drawn in separate blocks. For each step in the chain, a new b is drawn from a proposal density that depends on Σ, and that draw is accepted according to the MH rule applied to the target density $p(b|\Sigma,D)$. A new Σ is then drawn from a proposal density that depends on b and accepted according to the MH rule applied to the target density $p(\Sigma|b,D)$. The conditional density $p(b|\Sigma,D)$ is obtained by rewriting (A.27) and retaining factors involving only b:

$$p(b|\Sigma,D) \propto \exp\left\{-\frac{1}{2}(b-\hat{b})'(\Sigma^{-1}\otimes X'X)(b-\hat{b})\right\} \times(1-\rho^2)^{1/2}\exp\left\{-\frac{1-\rho^2}{2\sigma_v^2}\left(x_0-\frac{\theta}{1-\rho}\right)^2\right\},\ \rho\in(-1,\ 1). \quad \text{(A.28)}$$

The proposal density for b is specified as multivariate normal with mean $\hat{b}$ and covariance matrix $\Sigma\otimes(X'X)^{-1}$. In drawing Σ, it is more convenient to work with the conditional density of Σ^{-1} than Σ, and the Jacobian of that transformation is $|\Sigma|^3$. Multiplying (A.27) by that quantity, rewriting the result, and then retaining factors involving only Σ^{-1} gives

$$p(\Sigma^{-1}|b,D) \propto (\sigma^{11})^{-1/2}|\Sigma^{-1}|^{(T-2)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\,[S+(B-\hat{B})'X'X(B-\hat{B})]\Sigma^{-1}\right\} \times\exp\left\{-\frac{(1-\rho^2)|\Sigma^{-1}|}{2\sigma^{11}}\left(x_0-\frac{\theta}{1-\rho}\right)^2\right\}. \quad \text{(A.29)}$$

[19]For an introduction to the MH algorithm, see Chib and Greenberg (1995) or Gilks, Richardson, and Spiegelhalter (1996).

The relation $\sigma_v^2 = \sigma^{11}|\Sigma^{-1}|^{-1}$ is used in obtaining (A.29). The proposal density for Σ^{-1} is specified as Wishart with $T+1$ degrees of freedom and parameter matrix $[S + (B - \hat{B})'X'X(B - \hat{B})]^{-1}$. The derivations of the conditional densities from the joint density in (A.27) are aided by the observations that

$$\begin{aligned}(z - Zb)'(\Sigma^{-1} \otimes I_T)(z - Zb) & \\ &= (b - \hat{b})'(\Sigma^{-1} \otimes X'X)(b - \hat{b}) \ + \text{ terms without } b \qquad \text{(A.30)} \\ &= \text{tr}\,(Y - XB)'(Y - XB)\Sigma^{-1} \\ &= \text{tr}\,[S + (B - \hat{B})'X'X(B - \hat{B})]\Sigma^{-1}, \qquad \text{(A.31)}\end{aligned}$$

where (A.30) is used in obtaining (A.28), and (A.31) is used in obtaining (A.29).

With the likelihood in (25) and the prior in (27), the joint posterior density for b and Σ is given by

$$\begin{aligned}p(b, \Sigma|z, x_0) \ \propto \ & \sigma_v|\Sigma|^{-(T+5)/2} \exp\left\{-\frac{1}{2}(z - Zb)'(\Sigma^{-1} \otimes I_T)(z - Zb)\right\} \\ & \times(1 - \rho^2)^{-1/2} \exp\left\{-\frac{1 - \rho^2}{2\sigma_v^2}\left(x_0 - \frac{\theta}{1 - \rho}\right)^2\right\}. \ \rho \in (-1,\ 1). \qquad \text{(A.32)}\end{aligned}$$

Draws from the posteriors are again obtained using the MH algorithm. The proposal densities are identical to those used above, and the conditional densities are given by

$$\begin{aligned}p(b|\Sigma, z, x_0) \ \propto \ & \exp\left\{-\frac{1}{2}(b - \hat{b})'(\Sigma^{-1} \otimes X'X)(b - \hat{b})\right\} \\ & \times(1 - \rho^2)^{-1/2} \exp\left\{-\frac{1 - \rho^2}{2\sigma_v^2}\left(x_0 - \frac{\theta}{1 - \rho}\right)^2\right\}, \quad \rho \in (-1,\ 1), \qquad \text{(A.33)}\end{aligned}$$

and

$$\begin{aligned}p(\Sigma^{-1}|b, z, x_0) \ \propto \ & (\sigma^{11})^{1/2}|\Sigma^{-1}|^{(T-2)/2} \exp\left\{-\frac{1}{2}\text{tr}\,[S + (B - \hat{B})'X'X(B - \hat{B})]\Sigma^{-1}\right\} \\ & \times \exp\left\{-\frac{(1 - \rho^2)|\Sigma^{-1}|}{2\sigma^{11}}\left(x_0 - \frac{\theta}{1 - \rho}\right)^2\right\}. \qquad \text{(A.34)}\end{aligned}$$

All of the results reported based on the MH algorithms discussed above are based on 100,000 draws of b and Σ, obtained by retaining every 200th draw from a total of 20,000,000 draws after discarding an initial 40,000 "burn-in" draws. The acceptance rates for b range from 35% to 88%, depending on the sample period, while the acceptance rates for Σ are 94% or more. The 100,000 draws are used to compute the results in parts C and D of Table 2, and those same draws are used in generating draws from the predictive distribution of multiperiod returns (explained below).

A.7. Derivations of equations (28) and (31)

If b_2 and Σ are set equal to their MLE's, then maximizing (25) with respect to b_1 gives the MLE for that quantity. Observe from (20) that the value of b_1 that maximizes (25) must also minimize the term on the left-hand side of (A.30), given b_2 and Σ set equal to their MLE's. The first term on the right-hand side of (A.30) can be rewritten as

$$\begin{aligned}(b-\hat{b})'(\Sigma^{-1}\otimes X'X)(b-\hat{b}) \\ &= \left[b_1-\hat{b}_1-\frac{\sigma_{uv}}{\sigma_v^2}(b_2-\hat{b}_2)\right]'(\sigma^{11}X'X)\left[b_1-\hat{b}_1-\frac{\sigma_{uv}}{\sigma_v^2}(b_2-\hat{b}_2)\right] \\ &\quad + \text{terms without } b_1, \end{aligned} \tag{A.35}$$

using the relation $\sigma^{12}/\sigma^{11} = -\sigma_{uv}/\sigma_v^2$. The result in (31) then follows immediately.

From the joint posteriors for b and Σ in (A.27) and (A.32), the conditional mean of b_1 given b_2 and Σ is normal with mean

$$\mathrm{E}\{b_1|b_2,\Sigma,D\} = \hat{b}_1 + \frac{\sigma_{uv}}{\sigma_v^2}(b_2-\hat{b}_2), \tag{A.36}$$

which is obtained by again making use of (A.30) and (A.35). The same result obtains for the posterior obtained by combining the likelihood in (20) and the priors in (23) and (26), since the differences in the joint posteriors, involving b_2 and Σ, drop out in the conditional posterior for b_1. Equation (28) is the second row of the above vector equation (noting that, θ, the first element of b_2, does not enter the conditional mean for β).

A.8. Calculations involving the predictive distribution of $y_{T+K,(K)}$

Conditional on b, Σ, and x_T, it follows from (1), (3), (32), and the i.i.d. joint normality assumption for $(u_t\ v_t)$ that the distribution of $y_{T+K,(K)}$ is normal with mean and variance as follows. If $\rho \neq 1$, then the mean is given by

$$\mathrm{E}\{y_{T+K,(K)}|b,\Sigma,x_T\} = a_K + d_K x_T, \tag{A.37}$$

where

$$a_K = K\alpha + \beta\theta\left[\frac{K(1-\rho)-(1-\rho^K)}{(1-\rho)^2}\right] \tag{A.38}$$

and

$$d_K = \beta\frac{1-\rho^K}{1-\rho}. \tag{A.39}$$

The variance is given by

$$\begin{aligned}\text{var}\{y_{T+K,(K)}|b,\Sigma,x_T\} \\ &= K\sigma_u^2+\left(\frac{\beta}{1-\rho}\right)^2\left[K-1-\frac{2\rho(1-\rho^{K-1})}{1-\rho}+\frac{\rho^2(1-\rho^{2(K-1)})}{1-\rho^2}\right]\sigma_v^2 \\ &\quad +2\frac{\beta}{1-\rho}\left[K-1-\frac{\rho(1-\rho^{K-1})}{1-\rho}\right]\sigma_{uv}. \qquad \text{(A.40)}\end{aligned}$$

If $\rho = 1$ (an event with zero posterior measure), then

$$a_K = K\alpha+\beta\theta\frac{K(K-1)}{2}, \qquad \text{(A.41)}$$

$$d_K = K\beta, \qquad \text{(A.42)}$$

and

$$\text{var}\{y_{T+K,(K)}|b,\Sigma,x_T\} = K\sigma_u^2+(1/6)K(K-1)(2K-1)\beta^2\sigma_v^2+\beta K(K-1)\sigma_{uv}. \qquad \text{(A.43)}$$

In (38), $d = d_K$, $c = a_K + d\bar{x}$, and s equals the square root of the right-hand side of (A.40) or (A.43).

To compute the optimal stock allocation for a given K and x_T, 1 million draws from the predictive distribution of $y_{T+K,(K)}$ are generated by using (38) to draw 10 values of $y_{T+K,(K)}$ for each of the 100,000 values of b and Σ drawn from the posterior distribution (as explained previously). The average utility for these 1 million draws is computed for values of ω ranging from 0 to 1 in increments of .005, and the maximizing value of ω is reported in Table 3.

The moments in Tables 4 through 6 are computed using the 100,000 draws from the posterior distribution of b and Σ. The mean of $y_{T+K,(K)}$ is the average of the right-hand side of (A.37). The variance of $y_{T+K,(K)}$ is computed as the average of the right-hand side of (A.40) plus the variance of the right-hand side of (A.37). The third moment is computed by averaging the quantities appearing in the expectations on the right-hand side of (41).

A.9. Derivation of the Jeffreys prior in (52)

Since the conditional likelihood function remains in the same form as in the single-variable case, the earlier derivation of (27) requires only minor changes. The derivation proceeds virtually identically up to the first line of (A.23). That is

$$p(b,\Sigma^{-1}) \propto \left|\begin{array}{cc}\Sigma^{-1}\otimes\Psi & 0 \\ 0 & \frac{\partial^2\log|\Sigma|}{\partial\zeta\partial\zeta'}\end{array}\right|^{1/2}, \qquad \text{(A.44)}$$

except that Ψ in (A.20) becomes

$$\Psi = \frac{1}{T}\mathrm{E}\{X'X|b,\Sigma\} = \begin{bmatrix} 1 & \mu_h' \\ \mu_h & V_h + \mu_h\mu_h' \end{bmatrix}. \tag{A.45}$$

Proceeding as before, taking account of the fact that Σ is now $N\times N$ and Ψ is $(N+1)\times(N+1)$, gives

$$\begin{aligned} p(b,\Sigma^{-1}) &\propto \left|\Sigma^{-1}\otimes\Psi\right|^{1/2}\left|\frac{\partial^2 \log|\Sigma|}{\partial\zeta\partial\zeta'}\right|^{1/2} \\ &= \left(|\Sigma|^{-(N+1)}|\Psi|^N\right)^{1/2}\left(|\Sigma|^{N+1}\right)^{1/2} \\ &= |\Psi|^{N/2} \\ &= |V_h|^{N/2}. \end{aligned} \tag{A.46}$$

The last equality follows from the formula for the determinant of a partitioned matrix (e.g., Anderson, 1984, Theorem A.3.2). The Jacobian of the transformation from Σ^{-1} to Σ is $|\Sigma|^{-(N+1)}$, and multiplying (A.46) by that quantity gives (52). If (22) is applied separately for b and Σ, as discussed at the end of section A.5, then the prior for b is simply the first factor in the second line of (A.46), with $|\Sigma|^{-(N+1)/2}$ absorbed in the proportionality constant. In that case, the approximate Jeffreys prior is instead

$$p(b,\Sigma) \propto |V_h|^{N/2}|\Sigma|^{-(N+1)/2}. \tag{A.47}$$

Table 1

Finite-sample properties of $\hat{\beta}$

The table reports finite-sample properties of the ordinary least squares (OLS) estimator $\hat{\beta}$ in the regression,

$$y_t = \alpha + \beta x_{t-1} + u_t.$$

The sampling properties are computed under the assumption that x_t obeys the process

$$x_t = \theta + \rho x_{t-1} + v_t,$$

where $\rho^2 < 1$ and $[u_t \; v_t]'$ is distributed $N(0, \Sigma)$, identically and independently across t. The true bias and higher-order moments depend on ρ and Σ (with distinct elements σ_u^2, σ_v^2, and σ_{uv}). For each sample period, those parameters are set equal to the estimates obtained when y_t is the continuously compounded return in month t on the value-weighted NYSE portfolio, in excess of the one-month T-bill return, and x_t is the dividend-price ratio on the value-weighted NYSE portfolio at the end of month t. The moments in the standard setting are conditioned on $x_0, \ldots, x_{T-1}$. The p-values are associated with a test of $\beta = 0$ versus $\beta > 0$.

	Sample Period			
	1927–96	1927–51	1952–96	1977–96
A. True properties				
bias	0.07	0.18	0.18	0.42
standard deviation	0.16	0.33	0.27	0.45
skewness	0.71	0.83	0.98	1.29
kurtosis	3.84	4.14	4.62	5.83
p-value for $\beta = 0$	0.17	0.42	0.15	0.64
B. Properties in the standard regression setting				
bias	0	0	0	0
standard deviation	0.14	0.27	0.20	0.30
skewness	0	0	0	0
kurtosis	3.00	3.00	3.00	3.00
p-value for $\beta = 0$	0.06	0.22	0.02	0.26
C. Sample characteristics and parameter values				
$\hat{\beta}$	0.21	0.21	0.44	0.19
T	840	300	540	240
ρ	0.972	0.948	0.980	0.987
$\sigma_u^2 \times 10^4$	30.05	54.46	16.42	17.50
$\sigma_v^2 \times 10^4$	0.108	0.247	0.029	0.033
$\sigma_{uv} \times 10^4$	-1.621	-3.360	-0.651	-0.715

Table 2

Posterior distributions for β

The table reports Bayesian posterior moments for the slope coefficient in the regression,

$$y_t = \alpha + \beta x_{t-1} + u_t,$$

where y_t is the continuously compounded return in month t on the value-weighted NYSE portfolio, in excess of the one-month T-bill return, and x_t is the dividend-price ratio on the value-weighted NYSE portfolio at the end of month t. Also reported is the posterior probability that $\beta \leq 0$. It is assumed that x_t obeys the process

$$x_t = \theta + \rho x_{t-1} + v_t,$$

where $[u_t \; v_t]'$ is distributed $N(0, \Sigma)$, identically and independently across t. The method in Part A permits all elements of $b = (\alpha \; \beta \; \theta \; \rho)'$ to take values in the interval $(-\infty, \infty)$, whereas the methods in Parts B through D restrict ρ to the interval $(-1, 1)$. The methods in Part A and B are based on the "conditional" likelihood, which treats the initial observation x_0 as fixed. The methods in Parts B and C are based on the "exact" likelihood, which treats x_0 as a realization from its unconditional distribution.

	Sample Period			
	1927–96	1927–51	1952–96	1977–96
A. Conditional likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-\infty, \infty)$				
Mean	0.21	0.21	0.44	0.19
Std. Dev.	0.14	0.28	0.20	0.30
Skewness	0	0	0	0
Kurtosis	3.01	3.02	3.01	3.03
Prob.$\{\beta \leq 0\}$	0.06	0.22	0.02	0.26
B. Conditional likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-1, 1)$				
Mean	0.21	0.21	0.44	0.27
Std. Dev.	0.14	0.27	0.20	0.25
Skewness	0.02	0.04	0.12	0.45
Kurtosis	2.98	2.96	2.90	3.04
Prob.$\{\beta \leq 0\}$	0.06	0.22	0.01	0.13
C. Exact likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-1, 1)$				
Mean	0.23	0.26	0.38	0.38
Std. Dev.	0.14	0.26	0.18	0.24
Skewness	0.03	0.08	0.24	0.36
Kurtosis	2.97	2.95	2.93	3.02
Prob.$\{\beta \leq 0\}$	0.05	0.16	0.01	0.05
D. Exact likelihood; $p(b, \Sigma) \propto (1-\rho^2)^{-1}\sigma_v^2\lvert\Sigma\rvert^{-5/2}$, $\rho \in (-1, 1)$				
Mean	0.19	0.17	0.28	0.24
Std. Dev.	0.14	0.28	0.18	0.24
Skewness	0.00	0.04	0.37	0.53
Kurtosis	2.95	2.90	2.84	3.18
Prob.$\{\beta \leq 0\}$	0.10	0.28	0.05	0.16

Table 3

Optimal stock allocation (in percent) in a buy-and-hold strategy for various investment horizons and current dividend yields

The table reports optimal stock allocations implied by the predictive distribution for long-horizon returns. The investor is assumed to maximize the expected value of an iso-elastic utility function of terminal wealth with a coefficient of relative risk aversion is equal to 7. The predictive distribution is obtained using the two-equation model,

$$y_t = \alpha + \beta x_{t-1} + u_t$$
$$x_t = \theta + \rho x_{t-1} + v_t,$$

where y_t is the continuously compounded return in month t on the value-weighted NYSE portfolio, in excess of the one-month T-bill return, and x_t is the dividend-price ratio on the value-weighted NYSE portfolio at the end of month t. It is assumed $[u_t \; v_t]'$ is distributed $N(0, \Sigma)$, identically and independently across t. Define $b = (\alpha \; \beta \; \theta \; \rho)'$. The method in Parts A and D, based on the "conditional" likelihood, treat the initial observation x_0 as fixed. The methods in Parts B and C, based on the "exact" likelihood, restrict ρ to the interval $(-1, 1)$ and treat x_0 as a realization from its unconditional distribution.

	Current dividend yield (x_T)				
Investment horizon	2%	3%	4%	5%	6%
A. Conditional likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-\infty, \infty)$					
1 month	0	22	61	97	100
1 year	0	27	65	100	100
5 years	11	50	81	86	81
10 years	37	69	71	63	55
20 years	63	58	52	44	38
B. Exact likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-1, \; 1)$					
1 month	0	15	46	79	100
1 year	0	18	51	82	100
5 years	4	37	67	83	85
10 years	27	57	73	76	71
20 years	57	65	62	59	54
C. Exact likelihood; $p(b, \Sigma) \propto (1-\rho^2)^{-1}\sigma_v^2\lvert\Sigma\rvert^{-5/2}$, $\rho \in (-1, \; 1)$					
1 month	0	21	45	68	91
1 year	1	24	48	70	86
5 years	13	37	57	68	69
10 years	29	51	60	60	56
20 years	50	55	52	47	42
D. Conditional MLEs as true parameters (ignore estimation risk)					
1 month	0	22	60	98	100
1 year	0	27	68	100	100
5 years	7	55	100	100	100
10 years	45	92	100	100	100
20 years	100	100	100	100	100

Table 4

Expected excess return (in percent) for various investment horizons and current dividend yields

The table reports the mean of the predictive distribution for the long-horizon excess stock return, $y_{T+K,(K)} \equiv \sum_{k=1}^{K} y_{T+k}$, where $y_{t,(1)} = y_t$ and K is the length of the investment horizon (in months). The predictive distribution is obtained using the two-equation model,

$$y_t = \alpha + \beta x_{t-1} + u_t$$
$$x_t = \theta + \rho x_{t-1} + v_t,$$

where y_t is the continuously compounded return in month t on the value-weighted NYSE portfolio, in excess of the one-month T-bill return, and x_t is the dividend-price ratio on the value-weighted NYSE portfolio at the end of month t. It is assumed $[u_t \; v_t]'$ is distributed $N(0, \Sigma)$, identically and independently across t. Define $b = (\alpha \; \beta \; \theta \; \rho)'$. The method in Parts A and D, based on the "conditional" likelihood, treat the initial observation x_0 as fixed. The methods in Parts B and C, based on the "exact" likelihood, restrict ρ to the interval $(-1, 1)$ and treat x_0 as a realization from its unconditional distribution.

	Current dividend yield (x_T)				
Investment horizon	2%	3%	4%	5%	6%
A. Conditional likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-\infty, \infty)$					
1 month	-3.2	2.0	7.3	12.5	17.7
1 year	-2.3	2.4	7.0	11.6	16.2
5 years	0.4	3.2	6.1	9.0	11.9
10 years	1.9	3.7	5.6	7.4	9.3
20 years	3.1	4.1	5.2	6.2	7.2
B. Exact likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-1, 1)$					
1 month	-3.5	1.0	5.5	10.0	14.5
1 year	-2.7	1.4	5.4	9.4	13.5
5 years	-0.2	2.4	5.1	7.8	10.4
10 years	1.3	3.1	4.9	6.7	8.5
20 years	2.7	3.7	4.7	5.8	6.8
C. Exact likelihood; $p(b, \Sigma) \propto (1-\rho^2)^{-1}\sigma_v^2\lvert\Sigma\rvert^{-5/2}$, $\rho \in (-1, 1)$					
1 month	-1.4	1.9	5.2	8.5	11.9
1 year	-0.9	2.1	5.2	8.2	11.2
5 years	0.7	2.8	5.0	7.1	9.2
10 years	1.8	3.3	4.8	6.3	7.8
20 years	2.9	3.8	4.7	5.6	6.5
D. Conditional MLEs as true parameters (ignore estimation risk)					
1 month	-3.2	2.0	7.3	12.5	17.8
1 year	-2.4	2.3	7.0	11.7	16.4
5 years	0.0	3.1	6.1	9.2	12.3
10 years	1.6	3.6	5.6	7.6	9.6
20 years	2.9	4.0	5.1	6.2	7.3

Table 5

Standard deviation of the excess return (in percent) for various investment horizons and current dividend yields

The table reports the standard deviation of the predictive distribution for the long-horizon excess stock return, $y_{T+K,(K)} \equiv \sum_{k=1}^{K} y_{T+k}$, where $y_{t,(1)} = y_t$ and K is the length of the investment horizon (in months). The predictive distribution is obtained using the two-equation model,

$$y_t = \alpha + \beta x_{t-1} + u_t$$
$$x_t = \theta + \rho x_{t-1} + v_t,$$

where y_t is the continuously compounded return in month t on the value-weighted NYSE portfolio, in excess of the one-month T-bill return, and x_t is the dividend-price ratio on the value-weighted NYSE portfolio at the end of month t. It is assumed $[u_t \; v_t]'$ is distributed $N(0, \Sigma)$, identically and independently across t. Define $b = (\alpha \; \beta \; \theta \; \rho)'$. The method in Parts A and D, based on the "conditional" likelihood, treat the initial observation x_0 as fixed. The methods in Parts B and C, based on the "exact" likelihood, restrict ρ to the interval $(-1, 1)$ and treat x_0 as a realization from its unconditional distribution.

	Current dividend yield (x_T)				
Investment horizon	2%	3%	4%	5%	6%
A. Conditional likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-\infty, \infty)$					
1 month	14.1	14.1	14.1	14.1	14.2
1 year	13.5	13.1	13.0	13.2	13.7
5 years	11.4	10.7	10.5	10.7	11.4
10 years	10.2	9.6	9.3	9.4	9.9
20 years	10.8	9.7	9.1	9.2	9.9
B. Exact likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-1, 1)$					
1 month	14.2	14.2	14.1	14.2	14.2
1 year	13.6	13.3	13.2	13.4	13.8
5 years	11.4	10.9	10.8	11.1	11.7
10 years	9.9	9.6	9.5	9.7	10.2
20 years	8.9	8.7	8.7	8.8	9.1
C. Exact likelihood; $p(b, \Sigma) \propto (1-\rho^2)^{-1}\sigma_v^2 \lvert\Sigma\rvert^{-5/2}$, $\rho \in (-1, 1)$					
1 month	14.2	14.1	14.1	14.2	14.2
1 year	13.9	13.5	13.5	13.7	14.1
5 years	12.5	11.8	11.6	12.0	12.9
10 years	11.4	10.7	10.6	10.9	11.8
20 years	10.5	10.1	9.9	10.2	10.9
D. Conditional MLEs as true parameters (ignore estimation risk)					
1 month	14.0	14.0	14.0	14.0	14.0
1 year	12.8	12.8	12.8	12.8	12.8
5 years	9.6	9.6	9.6	9.6	9.6
10 years	7.8	7.8	7.8	7.8	7.8
20 years	6.5	6.5	6.5	6.5	6.5

Table 6

Skewness of the excess return for various investment horizons and current dividend yields

The table reports the skewness of the predictive distribution for the long-horizon excess stock return, $y_{T+K,(K)} \equiv \sum_{k=1}^{K} y_{T+k}$, where $y_{t,(1)} = y_t$ and K is the length of the investment horizon (in months). The predictive distribution is obtained using the two-equation model,

$$y_t = \alpha + \beta x_{t-1} + u_t$$
$$x_t = \theta + \rho x_{t-1} + v_t,$$

where y_t is the continuously compounded return in month t on the value-weighted NYSE portfolio, in excess of the one-month T-bill return, and x_t is the dividend-price ratio on the value-weighted NYSE portfolio at the end of month t. It is assumed $[u_t \; v_t]'$ is distributed $N(0, \Sigma)$, identically and independently across t. Define $b = (\alpha \; \beta \; \theta \; \rho)'$. The method in Parts A and D, based on the "conditional" likelihood, treat the initial observation x_0 as fixed. The methods in Parts B and C, based on the "exact" likelihood, restrict ρ to the interval $(-1, 1)$ and treat x_0 as a realization from its unconditional distribution.

	Current dividend yield (x_T)				
Investment horizon	2%	3%	4%	5%	6%
A. Conditional likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-\infty, \infty)$					
1 month	0.0	0.0	0.0	0.0	0.0
1 year	0.1	0.0	0.0	0.0	-0.1
5 years	0.4	0.2	0.0	-0.2	-0.4
10 years	0.8	0.5	0.1	-0.3	-0.7
20 years	6.1	3.1	0.6	-2.2	-6.1
B. Exact likelihood; $p(b, \Sigma) \propto \lvert\Sigma\rvert^{-3/2}$, $\rho \in (-1, 1)$					
1 month	0.0	0.0	0.0	0.0	0.0
1 year	0.0	0.0	0.0	0.0	-0.1
5 years	0.3	0.1	0.0	-0.2	-0.3
10 years	0.3	0.2	0.0	-0.2	-0.4
20 years	0.3	0.2	-0.1	-0.2	-0.4
C. Exact likelihood; $p(b, \Sigma) \propto (1-\rho^2)^{-1}\sigma_v^2 \lvert\Sigma\rvert^{-5/2}$, $\rho \in (-1, 1)$					
1 month	0.0	0.0	0.0	0.0	0.0
1 year	0.0	0.0	0.0	0.0	-0.1
5 years	0.3	0.1	0.0	-0.2	-0.4
10 years	0.5	0.2	-0.1	-0.4	-0.6
20 years	0.6	0.3	-0.1	-0.5	-0.8
D. Conditional MLEs as true parameters (ignore estimation risk)					
1 month	0.0	0.0	0.0	0.0	0.0
1 year	0.0	0.0	0.0	0.0	0.0
5 years	0.0	0.0	0.0	0.0	0.0
10 years	0.0	0.0	0.0	0.0	0.0
30 years	0.0	0.0	0.0	0.0	0.0

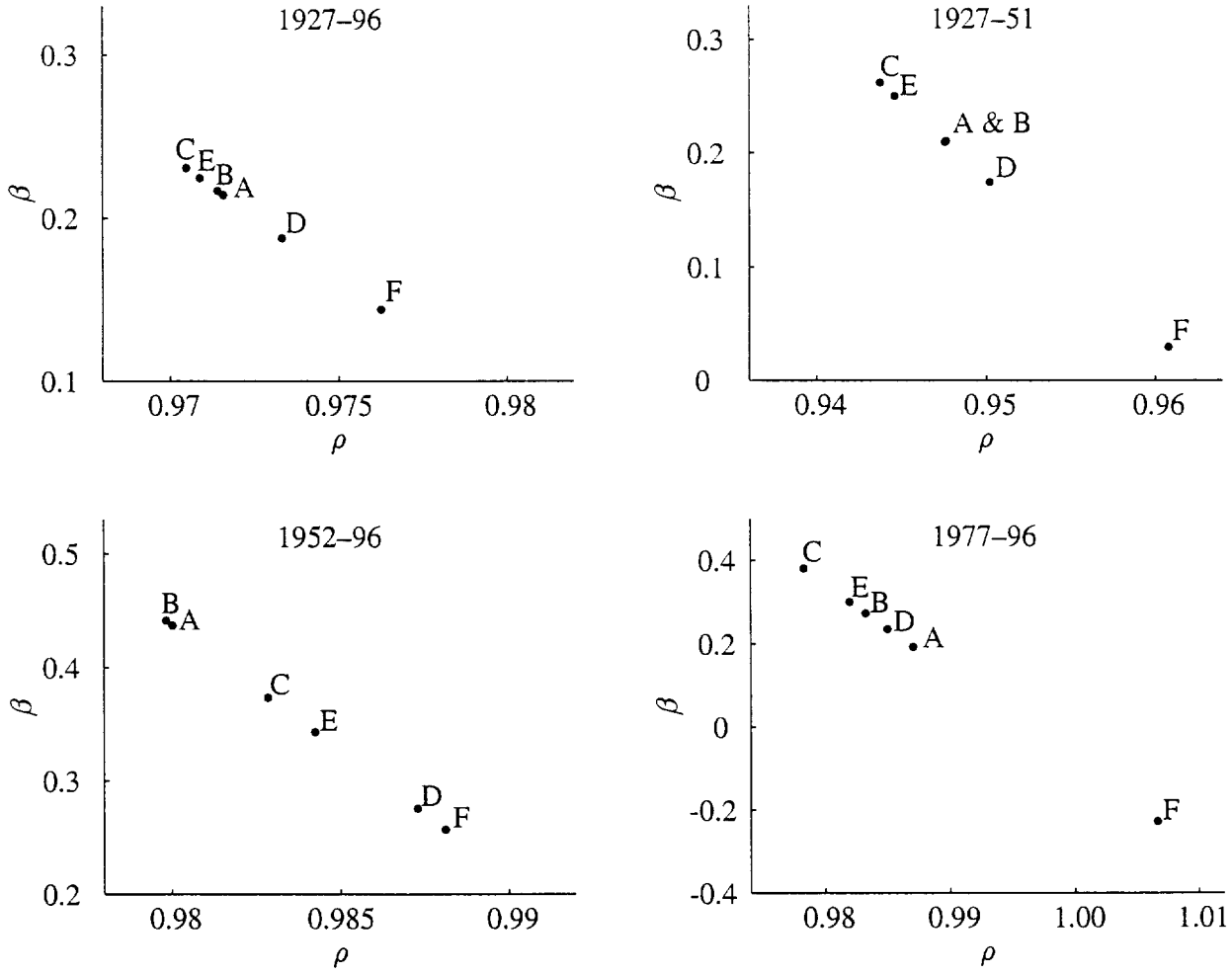

Figure 1. Estimates of β and ρ. The figure plots, for various methods and subperiods, the estimate of β versus the estimate of ρ for the two-equation model,

$$y_t = \alpha + \beta x_{t-1} + u_t$$
$$x_t = \theta + \rho x_{t-1} + v_t,$$

where y_t is the continuously compounded return in month t on the value-weighted NYSE portfolio, in excess of the one-month T-bill return, and x_t is the dividend-price ratio on the value-weighted NYSE portfolio at the end of month t. It is assumed $[u_t \; v_t]'$ is distributed $N(0, \Sigma)$, identically and independently across t. Define $b = (\alpha \; \beta \; \theta \; \rho)'$. The estimation methods are denoted as follows:

A: Bayesian posterior mean based on the conditional likelihood and $p(b, \Sigma) \propto |\Sigma|^{-3/2}$, $\rho \in (-\infty, \infty)$; also the ordinary least squares estimator; also the maximum-likelihood estimate (MLE) based on the conditional likelihood (treating x_0 as fixed).

B: Bayesian posterior mean based on the conditional likelihood and $p(b, \Sigma) \propto |\Sigma|^{-3/2}$, $\rho \in (-1, 1)$.

C: Bayesian posterior mean based on the exact likelihood (x_0 stochastic and drawn from its unconditional distribution) and $p(b, \Sigma) \propto |\Sigma|^{-3/2}$, $\rho \in (-1, 1)$.

D: Bayesian posterior mean based on the exact likelihood and $p(b, \Sigma) \propto (1-\rho^2)^{-1}\sigma_v^2|\Sigma|^{-5/2}$, $\rho \in (-1, 1)$.

E: MLE based on the exact likelihood (which assumes $\rho^2 < 1$ and x_0 is drawn from the unconditional distribution)

F: OLS estimates corrected for bias, where the bias is evaluated using ρ and Σ obtained from the OLS estimation.

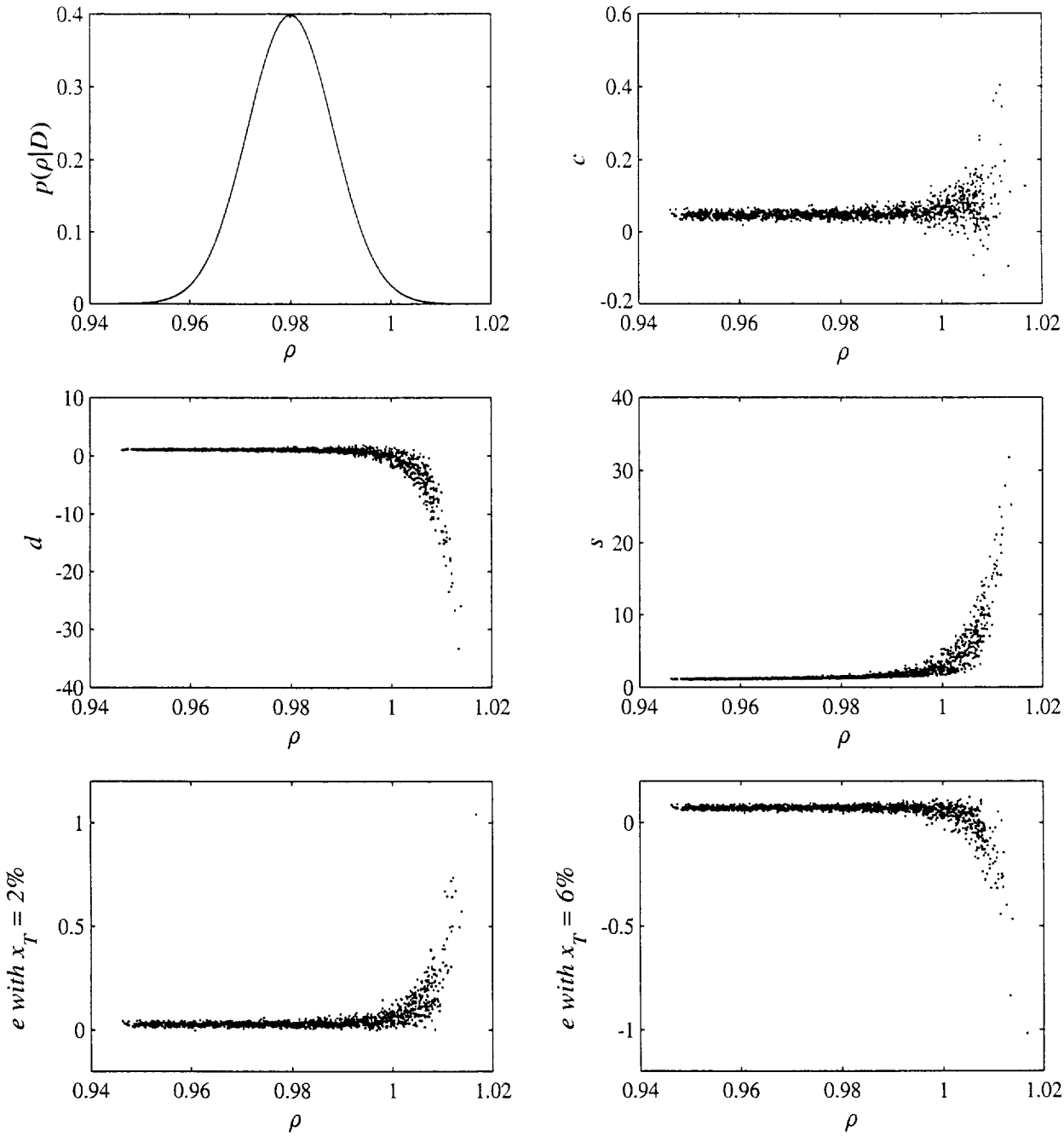

Figure 2. The role of ρ in the posterior distributions of long-horizon parameters. The upper left graph displays the posterior density of ρ, the slope coefficient in the relation $x_t = \theta + \rho x_{t-1} + v_t$, where x_t is the dividend yield in month t and v_t has zero mean conditional on x_{t-1}. The remaining five graphs display plots based on 100,000 draws from the joint posterior distribution of the model parameters. Each graph plots ρ versus one of the quantities in the relation

$$y_{T+K,(K)} = c + d(x_T - \bar{x}) + s\eta,$$

where η is a normal (0, 1) variate,

$$e = c + d(x_T - \bar{x}),$$

$y_{T+K,(K)}$ is the K-month excess return through month $T+K$, and $\bar{x}$ is the sample mean of x_t. The sample period is 1952–96, the return horizon is $K = 240$ (20 years), and the posteriors are obtained under the prior specification in which $p(\rho) \propto 1$, $\rho \in (-\infty, \infty)$. The return quantities are annualized, so that c, d, and e are multiplied by $(12/K)$ and s is multiplied by $(12/K)^{1/2}$. The scales are decimal values (not percents).

References

Anderson, T.W., 1984, An Introduction to Multivariate Statistical Analysis, Wiley, New York.

Barberis, N., 1999, Investing for the long run when returns are predictable, Journal of Finance, forthcoming.

Bauer, G.H., 1998, Currency hedging over short and long horizons with time-varying expected returns, working paper, University of Rochester, Rochester, NY.

Bekaert, G., Hodrick, R.J., Marshall, D.A., 1997, On biases in tests of the expectations hypothesis of the term structure of interest rates, Journal of Financial Economics 44, 309–348.

Berger, J.O., 1985, Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York.

Bernardo, J.M., Smith, A.F.M., 1994, Bayesian Theory, Wiley, Chichester.

Bilson, J.F.O., 1981, The 'speculative efficiency' hypothesis, Journal of Business 54, 435–452.

Box, G.E.P., Jenkins, G.M., 1970, Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, CA.

Box, G.E.P., Tiao, G.C., 1973, Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading, MA.

Campbell, J.Y., 1987, Stock returns and the term structure, Journal of Financial Economics 18, 373–399.

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

Chib, S., Greenberg, E., 1995, Understanding the Metropolis-Hastings algorithm, American Statistician 49, 327–335.

Elliott, G., Stock, J.H., 1994, Inference in time series regression when the order of integration of a regressor is unknown, Econometric Theory 10, 672–700.

Fama, E.F., 1984a, The information in the term structure, Journal of Financial Economics 13, 509–528.

Fama, E.F., 1984b, Forward and spot exchange rates, Journal of Monetary Economics 14, 319–338.

Fama. E.F., Bliss, R.R., 1987, The information in long-maturity forward rates, American Economic Review 77, 680–692.

Fama, E.F., Schwert, G.W., 1977, Asset returns and inflation, Journal of Financial Economics 5, 115–146.

Fama, E.F., French, K.R., 1988, Dividend yields and expected stock returns, Journal of Financial Economics 22, 3–26.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996, Markov Chain Monte Carlo in Practice, Chapman and Hall, London.

Hamilton, J.D., 1994, Time Series Analysis, Princeton University Press, Princeton, NJ.

Hastings, W.K., 1970, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57, 97–109.

Imhof, J.P., 1961, Computing the distribution of quadratic forms in normal variables, Biometrika 48, 419–426.

Jeffreys, H., 1961, Theory of Probability, Oxford University Press, Clarendon.

Kandel, S., Stambaugh, R.F., 1996, On the predictability of stock returns: An asset-allocation perspective, Journal of Finance 51, 385–424.

Keim, D.B., Stambaugh, R.F., 1986, Predicting returns in the stock and bond markets, Journal of Financial Economics 17, 357–390.

Kendall, M.G., 1954, Note on bias in the estimation of autocorrelation, Biometrika 41, 403–404.

Kendall, M., Sir, S.A., 1997, The Advanced Theory of Statistics, Vol. 1, Macmillan, New York.

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

Magnus, J.R., 1986, The exact moments of a ratio of quadratic forms in normal variables, Annales D'Économie et de Statistique 4, 95–109.

Mankiw, N.G., Shapiro, M., 1986, Do we reject too often: Small sample properties of tests of rational expectations models, Economics Letters 20, 139–145.

Mark, N.C., 1995, Exchange rates and fundamentals: Evidence on long-horizon predictability, American Economic Review 85, 201–218.

Leamer, E.E., 1991, Comment on 'To criticize the critics,' Journal of Applied Econometrics 6, 371–373.

Marriott, F.H.C., Pope, J.A., 1954, Bias in the estimation of autocorrelations, Biometrika 41, 390–402.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953, Equations of state calculations by fast computing machines, Journal of Chemical Physics 21, 1087–1092.

Nankervis, J.C., Savin, N. E., 1988, The exact moments of the least-squares estimator for the autoregressive model: Corrections and extensions, Journal of Econometrics 37, 318–388.

Nelson, C.R., Kim, M.J., 1993, Predictable stock returns: The role of small sample bias, Journal of Finance 48, 641–661.

Nicholls, D.F., Pope, A.L., 1988, Bias in the estimation of multivariate autoregressions, Australian Journal of Statistics 30A, 296–309.

Perks, F.J.A, 1947, Some observations on inverse probability including a new indifference rule, Journal of the Institute of Actuaries 73, 285–334.

Phillips, P.C.B., 1991, To criticize the critics: An objective Bayesian analysis of stochastic trends, Journal of Applied Econometrics 6, 333–364.

Pontiff, J., Schall, L.D., 1998, Book-to-market ratios as predictors of market returns, Journal of Financial Economics 49, 141–160.

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

Sawa, T., 1978, The exact moments of the least squares estimator for the autoregressive model, Journal of Econometrics 8, 159–172.

Shiller, R.J., Campbell, J.Y., Schoenholtz, K.L., 1983, Forward rates and future policy: Interpreting the term structure of interest rates, Brookings Papers on Economic Activity 1, 173–223.

Sims, C.A., 1988, Bayesian skepticism on unit root econometrics, Journal of Economic Dynamics and Control 12, 463–474.

Sims, C.A., Uhlig, H., 1991, Understanding unit rooters: A helicopter tour, Econometrica 59, 1591–1599.

Stambaugh, R.F., 1986, Bias in regressions with lagged stochastic regressors, working paper, University of Chicago, Chicago, IL.

Stock, J.H., 1991, Bayesian approaches to the 'unit root' problem: A comment, Journal of Applied Econometrics 6, 403–411.

Uhlig, H., 1994, On Jeffreys prior when using the exact likelihood function, Econometric Theory 10, 633–644.

Welch, B.L., Peers, H.W., 1963, On formulae for confidence points based on integrals of weighted likelihoods, Journal of the Royal Statistical Society, Series B 25, 318–329.

Zellner, A., 1971, An Introduction to Bayesian Inference in Econometrics, John Wiley and Sons, New York.