NBER WORKING PAPER SERIES

WHAT IS THE CHANCE THAT THE EQUITY PREMIUM VARIES OVER TIME? EVIDENCE FROM PREDICTIVE REGRESSIONS

Jessica A. Wachter
Missaka Warusawitharana

Working Paper 17334
http://www.nber.org/papers/w17334

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2011

We are grateful to Sean Campbell, Mark Fisher, Michael Johannes, Matthew Pritsker, Robert Stambaugh, Stijn van Nieuwerburgh, Jonathan Wright, Moto Yogo, Hao Zhou and seminar participants at the 2008 meetings of the American Finance Association, the 2007 CIRANO Financial Econometrics Conference, the 2007 Winter Meeting of the Econometric Society, the 2010 Federal Reserve Conference on Financial Markets, the Federal Reserve Board, the University of California at Berkeley and the Wharton School for helpful comments. We are grateful for financial support from the Aronson+Johnson+Ortiz fellowship through the Rodney L. White Center for Financial Research. This manuscript does not reflect the views of the Board of Governors of the Federal Reserve System or those of the National Bureau of Economic Research.

What is the Chance that the Equity Premium Varies over Time? Evidence from Predictive Regressions
Jessica A. Wachter and Missaka Warusawitharana
NBER Working Paper No. 17334
August 2011, Revised January 2012
JEL No. C11,C22,G11,G17

ABSTRACT

We examine the evidence on excess stock return predictability in a Bayesian setting in which the investor faces uncertainty about both the existence and strength of predictability. When we apply our methods to the dividend-price ratio, we find that even investors who are quite skeptical about the existence of predictability sharply modify their views in favor of predictability when confronted by the historical time series of returns and predictor variables. Correctly taking into account the stochastic properties of the regressor has a dramatic impact on inference, particularly over the 2000-2005 period.

Jessica A. Wachter
Department of Finance
2300 SH-DH
The Wharton School
University of Pennsylvania
3620 Locust Walk
Philadelphia, PA 19104
and NBER
jwachter@wharton.upenn.edu

Missaka Warusawitharana
Department of Research and Statistics
Board of Governors of the Federal Reserve
Mail Stop 97
20th and Constitution Ave
Washington D.C., 20551
Missaka.N.Warusawitharana@frb.gov

1 Introduction

In this study, we evaluate the evidence in favor of excess stock return predictability from the perspective of a Bayesian investor. We focus on the case of a single predictor variable to highlight the complex statistical issues that come into play in this deceptively simple problem.

The investor in our model considers the evidence in favor of the following linear model for excess returns:

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

where r_{t+1} denotes the return on a broad stock index in excess of the riskfree rate, x_t denotes a predictor variable, and u_{t+1} the unpredictable component of the return. The investor also places a finite probability on the following model:

$$r_{t+1} = \alpha + u_{t+1}. \tag{2}$$

Namely, the investor assigns a prior probability q to the state of the world in which returns are predictable (because the prior on β will be smooth, the chance of $\beta = 0$ in (1) is infinitesimal), and a probability $1-q$ to the state of the world in which returns are completely unpredictable. In both cases, the parameters are unknown. Thus our model allows for both parameter uncertainty and "model uncertainty".[1]

Allowing for a non-zero probability on (2) is one way in which we depart from previous studies. Previous Bayesian studies of return predictability allow for uncertainty in the parameters in (1), but assume flat priors (see Barberis (2000), Brandt, Goyal, Santa-Clara, and Stroud (2005), Johannes, Polson, and Stroud (2002), Skoulakis (2007) and Stambaugh (1999)). As Wachter (2010) shows, flat or nearly-flat priors imply a degree of predictability that is hard to justify economically. Other studies (Kandel and Stambaugh (1996), Pastor and Stambaugh (2009), Shanken and Tamayo (2011), Wachter and Warusawitharana (2009))

[1] However, note that our investor is Bayesian, rather than ambiguity averse (e.g. Chen and Epstein (2002)).Our priors are equivalent to placing a point mass on $\beta = 0$ in (1).

investigate the impact of economically informed prior beliefs. These studies nonetheless assume that the investor places a probability of one on the predictability of returns. However, an investor who thinks that (2) represents a compelling null hypothesis will have a prior that places some weight on the possibility that returns are not predictable at all.

Our work also relates to the Bayesian model selection methods of Avramov (2002) and Cremers (2002). In these studies, the investor has a prior probability over the full set of possible linear models that make use of a given set of predictor variables. Thus the setting of these papers is more complex than ours in that many predictor variables are considered. However, these papers also make the assumption that the predictor variables are either non-stochastic, or that their shocks are uncorrelated with shocks to returns. These assumptions are frequently satisfied in a standard ordinary least squares regression, but rarely satisfied in a predictive regression. In contrast, we are able to formulate and solve the Bayesian investor's problem when the regressor is stochastic and correlated with returns.

When we apply our methods to the dividend-price ratio, we find that an investor who believes that there is a 50% probability of predictability prior to seeing the data updates to a 86% posterior probability after viewing quarterly postwar data. We find average certainty equivalent returns of 1% per year for an investor whose prior probability in favor of predictability is just 20%. For an investor who believes that there is a 50/50 chance of return predictability, certainty equivalent returns are 1.72%.

We also empirically evaluate the effect of correctly incorporating the initial observation of the dividend-price ratio into the likelihood (the exact likelihood approach) versus the more common conditional likelihood approach. In the conditional likelihood approach, the initial observation of the predictor variable is treated as a known parameter rather than as a draw from the data generating process. We find that the the unconditional risk premium is poorly estimated when we condition on the first observation. However, when this is treated as a draw from the data generating process, the expected return is estimated reliably. Surprisingly, the posterior mean of the unconditional risk premium differs from the sample average.

Finally, when we examine the evolution of posterior beliefs over the postwar period, we find substantial differences between the beliefs implied by our approach, which treats the regressor as stochastic and realistically captures the relation between the regressor and returns, and beliefs implied by assuming non-stochastic regressors. In particular, our approach implies that the belief in the predictability of returns rises dramatically over the 2000-2005 period while approaches assuming fixed regressors imply a decline.

The remainder of the paper is organized as follows. Section 2 describes our statistical method and contrasts it with alternative approaches. Section 3 describes our empirical results. Section 4 concludes.

2 Statistical Method

2.1 Data generating processes

Let r_{t+1} denote continuously compounded excess returns on a stock index from time t to $t+1$ and x_t the value of a (scalar) predictor variable. We assume that this predictor variable follows the process

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

Stock returns can be predictable, in which case they follow the process (1) or unpredictable, in which case they follow the process (2). In either case, errors are serially uncorrelated, homoskedastic, and jointly normal:

$$\begin{bmatrix} u_{t+1} \\ v_{t+1} \end{bmatrix} \,|\, r_t, \ldots, r_1, x_t, \ldots, x_0 \sim N(0, \Sigma), \tag{4}$$

and

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}. \tag{5}$$

As we show below, the correlation between innovations to returns and innovations to the predictor variable implies that (3) affects inference about returns, even when there is no predictability.

When the process (3) is stationary, i.e. ρ is between -1 and 1, the predictor variable has an unconditional mean of

$$\mu_x = \frac{\theta}{1-\rho} \tag{6}$$

and a variance of

$$\sigma_x^2 = \frac{\sigma_v^2}{1-\rho^2}. \tag{7}$$

These follow from taking unconditional means and variances on either side of (3). Note that these are population values conditional on knowing the parameters. Given these, the population R^2 is defined as

$$\text{Population } R^2 = \frac{\beta^2\sigma_x^2}{\beta^2\sigma_x^2 + \sigma_u^2}.$$

2.2 Prior Beliefs

The investor faces uncertainty both about the model (i.e. whether returns are predictable or not), and about the parameters of the model. We represent this uncertainty through a hierarchical prior. There is a probability q, that investors face the distribution given by (1), (3) and (4). We denote this state of the world H_1. There is a probability $1-q$ that investors face the distribution given by (2), (3) and (4). We denote this state of the world H_0. As we will show, the stochastic properties of x have relevance in both cases.

The prior information on the parameters is conditional on H_i. Let

$$b_0 = [\alpha,\ \theta,\ \rho]^\top$$

and

$$b_1 = [\alpha,\ \beta,\ \theta, \rho]^\top.$$

Note that $p(b_1, \Sigma|H_1)$ can also be written as $p(\beta, b_0, \Sigma|H_1)$.[2] We set the prior on b_0 and Σ so that

$$p(b_0, \Sigma|H_0) = p(b_0, \Sigma|H_1) = p(b_0, \Sigma).$$

[2] Formally we could write down $p(b_1, \Sigma|H_0)$ by assuming $p(\beta|b_0, \Sigma, H_0)$ is a point mass at zero.

We assume the investor has uninformative beliefs on these parameters. We follow the approach of Stambaugh (1999) and Zellner (1996), and derive a limiting Jeffreys prior as explained in Appendix A. As Appendix A shows, this limiting prior takes the form

$$p(b_0, \Sigma) \propto \sigma_x \sigma_u |\Sigma|^{-\frac{5}{2}}, \tag{8}$$

for $\rho \in (-1, 1)$, and zero otherwise.

The parameter that distinguishes H_0 from H_1 is β. One approach would be to write down a prior distribution for β unconditional on the remaining parameters. However, it is difficult to think about priors on β in isolation from beliefs about other parameters. For example, a high variance of x_t might lower one's prior on β, while a large residual variance of r_t might raise it. Rather than placing a prior on β directly, we follow Wachter and Warusawitharana (2009) and place a prior on the population R^2. To implement this prior on the R^2, we place a prior on "normalized" β, that is β adjusted for the variance of x and the variance of u. Let

$$\eta = \sigma_u^{-1} \sigma_x \beta.$$

denote normalized β. We assume that prior beliefs on η are given by

$$\eta | H_1 \sim N(0, \sigma_\eta^2) \tag{9}$$

The population R^2 is closely related to η:

$$\text{Population } R^2 = \frac{\beta^2 \sigma_x^2}{\beta^2 \sigma_x^2 + \sigma_u^2} = \frac{\eta^2}{\eta^2 + 1}. \tag{10}$$

Equation (10) provides a mapping between a prior distribution on η and a prior distribution on the population R^2. Given an η draw, an R^2 draw can be computed using (10).

A prior on η implies a hierarchical prior on β. The prior for η, (9), implies

$$\beta | \alpha, \theta, \rho, \Sigma \sim N(0, \sigma_\beta^2), \tag{11}$$

where

$$\sigma_\beta = \sigma_\eta \sigma_x^{-1} \sigma_u.$$

Because σ_x is a function of ρ and σ_v, the prior on β is also implicitly a function of these parameters. The parameter σ_η indexes the degree to which the prior is informative. As $\sigma_\eta \to \infty$, the prior over β becomes uninformative; all values of β are viewed as equally likely. As $\sigma_\eta \to 0$, the prior converges to $p(b_0, \Sigma)$ multiplied by a point mass at 0, implying a dogmatic view in no predictability. Combining (11) with (8) implies the joint prior under H_1:

$$
\begin{aligned}
p(b_1, \Sigma | H_1) &= p(\beta | b_0, \Sigma, H_1) p(b_0 | H_1) \\
&\propto \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \sigma_x^2 |\Sigma|^{-\frac{5}{2}} \exp\left\{ -\frac{1}{2} \beta^2 \left(\sigma_\eta^2 \sigma_x^{-2} \sigma_u^2 \right)^{-1} \right\}. \qquad (12)
\end{aligned}
$$

Jeffreys invariance theory provides an independent justification for modeling priors on β as (11). Stambaugh (1999) shows that the limiting Jeffreys prior for b_1 and Σ equals

$$
p(b_1, \Sigma | H_1) \propto \sigma_x^2 \, |\Sigma|^{-\frac{5}{2}}. \qquad (13)
$$

This prior corresponds to the limit of (12) as σ_η approaches infinity. Modeling the prior for β as depending on σ_x not only has a convenient interpretation in terms of the distribution of the R^2, but also implies that an infinite prior variance represents ignorance as defined by Jeffreys (1961). Note that a prior on β that is independent of σ_x would not have this property.

Figure 1 shows the resulting distribution for the population R^2 for various values of σ_η. Panel A shows the distribution conditional on H_1 while Panel B shows the unconditional distribution. More precisely, for any value k, Panel A shows the prior probability that the R^2 exceeds k, conditional on the existence of predictability. For large values of σ_η, e.g. 100, the prior probability that the R^2 exceeds k across the relevant range of values for the R^2 is close to one. The lower the value of σ_η, the less variability in β around its mean of zero, and the lower the probability that the R^2 exceeds k for any value of k. Panel B shows the unconditional probability that the R^2 exceeds k for any value of k, assuming that the prior probability of predictability, q, is equal to 0.5. By the definition of conditional probability:

$$
p(R^2 > k) = p(R^2 > k | H_1) q.
$$

Therefore Panel B takes the values in Panel A and scales them down by 0.5.

2.3 Likelihood

2.3.1 Likelihood under H_1

Under H_1, returns and the predictor variable follow the joint process given in (1) and (3). It is convenient to group observations on returns and contemporaneous observations on the state variable into a matrix Y and lagged observations on the state variable and the constant into a matrix X. Let

$$Y = \begin{bmatrix} r_1 & x_1 \\ \vdots & \vdots \\ r_T & x_T \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_0 \\ \vdots & \vdots \\ 1 & x_{T-1} \end{bmatrix},$$

and let

$$\begin{aligned} z &= \text{vec}(Y) \\ Z_1 &= I_2 \otimes X. \end{aligned}$$

In the above, the vec operator stacks the elements of the matrix columnwise. It follows that the likelihood conditional on H_1 and on the first observation x_0 takes the form of

$$p(D|b_1, \Sigma, x_0, H_1) = |2\pi\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}(z - Z_1 b_1)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_1 b_1)\right\} \tag{14}$$

(see Zellner (1996)).

The likelihood function (14) conditions on the first observation of the predictor variable, x_0. Stambaugh (1999) argues for treating x_0 and $x_1, \ldots, x_T$ symmetrically: as random draws from the data generating process. If the process for x_t is stationary and has run for a substantial period of time, then results in Hamilton (1994, p. 265) imply that x_0 is a draw from a multivariate normal distribution with mean μ_x and standard deviation σ_x. Combining the likelihood of the first observation with the likelihood of the remaining T observations produces

$$p(D|b_1, \Sigma, H_1) = |2\pi\sigma_x^2|^{-\frac{1}{2}} |2\pi\Sigma|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \left(x_0 - \mu_x\right)^2 \sigma_x^{-2} - \frac{1}{2}(z - Z_1 b_1)^\top \left(\Sigma^{-1} \otimes I_T\right) (z - Z_1 b_1) \right\}. \quad (15)$$

Following Box and Tiao (1973), we refer to (14) as the *conditional likelihood* and (15) as the *exact likelihood.*

2.3.2 Likelihood under H_0

Under H_0, returns and the predictor variable follow the processes given in (2) and (3). Let

$$Z_0 = \begin{bmatrix} \iota_T & 0_{T\times 2} \\ 0_{T\times 1} & X \end{bmatrix},$$

where ι_T is the $T \times 1$ vector of ones. Then the conditional likelihood can be written as

$$p(D|b_0, \Sigma, x_0, H_0) = |2\pi\Sigma|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2}(z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right) (z - Z_0 b_0) \right\}. \quad (16)$$

The conditional likelihood takes the same form as in the seemingly unrelated regression model (see Ando and Zellner (2010)). Using similar reasoning as in the H_1 case, the exact likelihood is given by

$$p(D|b_0, \Sigma, H_0) = |2\pi\sigma_x^2|^{-\frac{1}{2}} |2\pi\Sigma|^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \left(x_0 - \mu_x\right)^2 \sigma_x^{-2} - \frac{1}{2}(z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right) (z - Z_0 b_0) \right\}. \quad (17)$$

As above, we refer to (16) as the *conditional likelihood* and (17) as the *exact likelihood.*

2.4 Posterior distribution

The investor updates his prior beliefs to form the posterior distribution upon seeing the data. As we discuss below, this posterior requires the computation of two quantities: the posterior of the parameters conditional on the absence or presence of return predictability, and the posterior probability that returns are predictable. Given these two quantities, we can simulate from the posterior distribution.

To compute the posteriors, we apply Bayes' rule conditional on the model:

$$p(b_i, \Sigma | H_i, D) \propto p(D | b_i, \Sigma, H_i) p(b_i, \Sigma | H_i), \quad i = 0, 1. \tag{18}$$

Because σ_x is a nonlinear function of the underlying parameters, the posterior distributions conditional on H_0 and H_1 are nonstandard and must by computed numerically. We can sample from these distributions quickly and accurately using the Metropolis-Hastings algorithm (see Chib and Greenberg (1995), Johannes and Polson (2006)). See Appendix B for details.

Let $\bar{q}$ denote the posterior probability that excess returns are predictable. By definition,

$$\bar{q} = p(H_1 | D).$$

It follows from Bayes' rule, that

$$\bar{q} = \frac{\mathcal{B}_{10} q}{\mathcal{B}_{10} q + (1 - q)}, \tag{19}$$

where

$$\mathcal{B}_{10} = \frac{p(D | H_1)}{p(D | H_0)} \tag{20}$$

is the Bayes factor for the alternative hypothesis of predictability against the null of no predictability. The Bayes factor is a likelihood ratio in that it is the likelihood of return predictability divided by the likelihood of no predictability. However, it differs from the standard likelihood ratio in that the likelihoods $p(D|H_i)$ are not conditional on the values of the parameters. These likelihoods are given by

$$p(D | H_i) = \int p(D | b_i, \Sigma, H_i) p(b_i, \Sigma | H_i) \, db_i \, d\Sigma, \qquad i = 0, 1. \tag{21}$$

To form these likelihoods, the likelihoods conditional on parameters (the likelihood functions generally used in classical statistics) are integrated over the prior distribution of the parameters. Under our distributions, these integrals cannot be computed analytically. However, the Bayes factor (20) can be computed directly using the generalized Savage-Dickey ratio (Dickey (1971), Verdinelli and Wasserman (1995)). Details can be found in Appendix C.

Putting these two pieces together, we draw from the posterior parameter distribution by drawing from $p(b_1, \Sigma | D, H_1)$ with probability $\bar{q}$ and from $p(b_0, \Sigma | D, H_0)$ with probability $1 - \bar{q}$.

2.5 An alternative: Non-stochastic regressors

An alternative approach to inference is to adopt the standard assumptions of ordinary least squares regression, namely that the regressors x_t are fixed, or that u_s and v_t are uncorrelated for all s and t. For example, consider the priors and likelihood proposed by Fernandez, Ley, and Steel (2001). Let $\hat{\sigma}_x$ denote the sample variance of x:

$$\hat{\sigma}_x = \frac{1}{T}\sum_{t=1}^{T}\left(x_t - \frac{1}{T}\sum_{s=1}^{T} x_s\right)^2.$$

Fernandez et al. propose the following priors on α, β and σ_u:

$$p(\beta \,|\, \sigma_u^2, H_1) = N(0, \kappa\sigma_u^2\hat{\sigma}_x^{-1}), \tag{22}$$

where κ is a constant that determines the informativeness of the prior, and

$$p(\sigma_u) \propto \sigma_u^{-1}. \tag{23}$$

The specification is completed by setting $p(\alpha) \propto 1$. These assumptions on the prior are combined with the likelihood

$$p(D \,|\, \alpha, \beta, \sigma_u, H_1) = \left(2\pi\sigma_u^2\right)^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\sum_{t=0}^{T-1}(r_{t+1} - \alpha - \beta x_t)^2\sigma_u^{-2}\right\} \tag{24}$$

and

$$p(D \,|\, \alpha, \beta, \sigma_u, H_0) = \left(2\pi\sigma_u^2\right)^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\sum_{t=0}^{T-1}(r_{t+1} - \alpha)^2\sigma_u^{-2}\right\}. \tag{25}$$

The expressions for the prior and likelihood under H_0 are analogous. Similar specifications are employed by Chipman, George, and McCulloch (2001), Cremers (2002), Stock and Watson (2011) and Wright (2008). This formulation is closely related to the conjugate prior described in Zellner (1996). Like Zellner's prior, it leads to analytical expressions for the posterior distribution and Bayes factor.

The above specification differs from ours in two ways. First, the prior beliefs condition on observations on x_t. Second, and more fundamentally, the likelihood function conditions on x_t, namely, it treats it as known. The two are related, in that both assumptions are most reasonable in cases where x_t is known at time zero, or where all correlations between u_t and v_t are zero, but not otherwise. To see this, note that (24) takes the form

$$\prod_{t=0}^{T-1} p(r_{t+1}|x_t, \alpha, \beta, \sigma_u) \tag{26}$$

where the terms in the product are given by the normal density. However, the true conditional likelihood is the product of multivariate normal terms:[3]

$$\prod_{t=0}^{T-1} p(r_{t+1}, x_{t+1}|x_t, b_1, \Sigma).$$

One could separate out the terms in the product as follows

$$\prod_{t=0}^{T-1} p(r_{t+1}|x_t, \alpha, \beta, \sigma_u)p(x_{t+1}|r_{t+1}, x_t, b_1, \Sigma). \tag{27}$$

Under the assumption that u_t and v_t are uncorrelated in the prior, the second term will not depend on α, β and σ_u, and thus it "drops out" when multiplied by the prior to form the posterior in Bayes rule (18).[4] However, requiring these shocks to be uncorrelated is not realistic. For this reason, is it not generally valid to drop the second term in (27).

Perhaps there is some other way to justify the use of (24) rather than the full likelihood. Consider, for example, the following tempting (but wrong) argument. For convenience, define the notation $D_r = \{r_1, \ldots, r_T\}$ and $D_x = \{x_0, x_1, \ldots, x_T\}$. The marginal likelihood for returns (24) is valid regardless of the assumption on the correlation between u_t and v_t. Could one form a marginal posterior, $p(\alpha, \beta, \sigma_u|D_r)$? One way to do this might be to consider

$$p(\alpha, \beta, \sigma_u|D_r) \propto p(D_r|\alpha, \beta, \sigma_u)p(\alpha, \beta, \sigma_u).$$

However, $p(\alpha, \beta, \sigma_u|D_r)$ is not the same as (24), which uses information on x as well as r.

[3]Note that this expression still omits the likelihood for the initial observation x_0.

[4]More precisely, it becomes part of the constant term.

Another tempting but incorrect argument is to form the posterior

$$p(\alpha, \beta, \sigma_u | D_r, D_x) \propto p(D_r|\alpha, \beta, \sigma_u, D_x) p(\alpha, \beta, \sigma_u),$$

which represents valid inference under the assumption $p(\alpha, \beta, \sigma_u) = p(\alpha, \beta, \sigma_u | D_x)$. Again, however, $p(D_r|\alpha, \beta, \sigma_u, D_x)$ does not actually represent the likelihood (24). In this case, the reason is

$$p(D_r|\alpha, \beta, \sigma_u, D_x) = \prod_{t=0}^{T-1} p(r_{t+1}|x_t, \alpha, \beta, \sigma_u, D_x).$$

Instead of just computing r_{t+1} knowing x_t, $p(r_{t+1}|x_t, \alpha, \beta, \sigma_u, D_x)$ requires one to computes the likelihood of each observation r_{t+1} knowing the full time series of x. Because of the correlation between u and v, future shocks to x convey additional information about returns. While technically speaking this approach is valid, it makes very little economic sense (why would x be observed before r?) and in any case is not implemented in any of the studies cited above.

Because x_t appears in the likelihood function, it cannot be simply ignored. Nor can it be treated as known. The only alternative is to assume that it is stochastic, as we have done. At the root of the problem is the fact that the similarity between the likelihood in the linear regression model in the time series setting and under OLS is only apparent. In a time series setting, it is not valid to condition on the entire time path of the "independent" variable. The differences ultimately come down to the interpretation of the shock u_t. In a standard OLS setting, u_t is an error, and is thus correlated with the independent variable at all leads and lags. In a time series setting, it is not an error, but rather a shock, and this independence does not hold.

3 Results

3.1 Data

We use data from the Center for Research on Security Prices (CRSP). We compute excess stock returns by subtracting the continuously compounded 3-month Treasury bill return from the return on the value-weighted CRSP index at a quarterly frequency. Following a large empirical literature on return predictability, we focus on the dividend-price ratio as the regressor because the present-value relation between prices and returns suggests that it should capture variables that predict stock returns. The dividend-price ratio is computed by dividing the dividend payout over the previous 12 months with the current price of the stock index. The use of 12 months of data accounts for seasonalities in dividend payments. We use the logarithm of the dividend-price ratio as the predictor variable. Data are quarterly from 1952 to 2009.

3.2 Bayes factors and posterior means

Table 1 reports Bayes factors for various choices on the prior distribution. Four values of σ_η are considered: 0.051, 0.087, 0.148 and 100. These translate into values of $P(R^2 > .01|H_1)$ (the prior probability that the R^2 exceeds 0.01) equal to 0.05, 0.25, 0.50 and 0.99 respectively. These R^2s should be interpreted in terms of regressions performed at a quarterly frequency. Bayes factors are reported for the exact likelihood, and, to evaluate the importance of including the initial term, the conditional likelihood as well.

Table 1 shows that the Bayes factor is hump-shaped in $P(R^2 > 0.01|H_1)$. For small values, the Bayes factor is close to one. For large values, the Bayes factor is close to zero. Both results can be understood using the formula for the Bayes factor in (20) and for the likelihoods $p(D \mid H_i)$ in (21). For low values of this probability, the investor imposes a very tight prior on the R^2. Therefore the hypotheses that returns are predictable and that returns are unpredictable are nearly the same. It follows from (21) that the likelihoods of the data under these two scenarios are nearly the same and that the Bayes factor is nearly one. This is intuitive: when two hypotheses are close, a great deal of data are required to distinguish one from the other.

The fact that the Bayes factor approaches zero as $P(R^2 >, 01|H_1)$ increases is less intuitive. The reduction in Bayes factors implies that, as the investor allows a greater range of values for the R^2, the posterior probability that returns are predictable approaches zero. This effect is known as Bartlett's paradox, and was first noted by Bartlett (1957) in the context of distinguishing between uniform distributions. As Kass and Raftery (1995) discuss, Bartlett's paradox makes it crucial to formulate an informative prior on the parameters that differ between H_0 and H_1. The mathematics leading to Bartlett's paradox are most easily seen in a case where Bayes factors can be computed in closed form. However, we can obtain an understanding of the paradox based on the form of the likelihoods $p(D \mid H_1)$ and $P(D \mid H_0)$. These likelihoods involve integrating out the parameters using the prior distribution. If the prior distribution on β is highly uninformative, the prior places a large amount of mass in extreme regions of the parameter space. In these regions, the likelihood of the data conditional on the parameters will be quite small. At the same time, the prior places a relatively small amount of mass in the regions of the parameter space where the likelihood of the data is large. Therefore $P(D \mid H_1)$ (the integral of the likelihood under H_1) is small relative to $P(D \mid H_0)$ (the integral of the likelihood under H_0).

Table 1 also shows that there are substantial differences between the Bayes factors resulting from the exact versus the conditional likelihood.[5] The Bayes factors resulting from the exact likelihood are larger than those resulting from the conditional likelihood, thus implying a greater posterior probability of return predictability. This difference reflects the fact that the posterior mean of β, conditional on H_1, is higher for the exact likelihood than for the conditional likelihood, and the posterior mean is ρ is lower.[6]

We can use the connection between the posterior means and the Bayes factor to under-

[5]We are not the first to note the importance of the first observation. See, for example, Poirier (1978).

[6]The source of this negative relation is the negative correlation between shocks to returns and shocks to the predictor variable. Suppose that a draw of β is below its value predicted by ordinary least squares (OLS). This implies that the OLS value for β is "too high", i.e. in the sample shocks to the predictor variable are followed by shocks to returns of the same sign. Therefore shocks to the predictor variable tend to be followed by shocks to the predictor variable that are of different signs. Thus the OLS value for ρ is "too low". This explains why values of the posterior mean of ρ are higher for low values of $P(R^2 > 0.01|H_1)$ (and hence low values of the posterior mean of β) than for high values, and higher than the ordinary least squares estimate.

stand why the Bayes factor changes with the specification. Using the exact likelihood leads to lower posterior values of ρ. This is because the exact likelihood leads to more precise estimates of μ_x, and therefore lower estimates of σ_x. Because the posterior mean of ρ is lower, the posterior mean of β is higher, and the Bayes factor is higher.

3.3 The long-run equity premium

For the predictability model, the expected excess return on stocks (the equity premium) varies over time. In the long run, however, the current value of x_t becomes irrelevant. Under our assumptions x_t is stationary with mean μ_x, and therefore r_t is also stationary with mean

$$\mu_r = E[\alpha + \beta x_t + u_{t+1} | b_1, \Sigma] = \alpha + \beta \mu_x.$$

As is the case with μ_x, this is a population value that conditions on the value of the parameters. For the no-predictability model, μ_r is simply equal to α. We can think of μ_r as the average equity premium; the fact that it is "too high" constitutes the equity premium puzzle (Mehra and Prescott (1985)), and it is often computed by simply taking the sample average of excess returns.

The posterior expectation of μ_r under various specifications is shown in the fifth column of Table 1. Because differences in the expected return arise from differences in the posterior mean of the predictor variable x, the table also reports the posterior mean of μ_x. The differences in the long-run equity premium are striking. The sample average of the (continuously compounded) excess return on stocks over this period is 4.49%. However, assuming the exact likelihood implies produces a range for this excess return between 3.45% and 3.90% depending on the strength of the prior. Why is the equity premium in these cases as much as a full percentage point lower?

To answer this question, it is helpful to look at the posterior means of the predictor variable, reported in the next column of Table 1. For the exact likelihood specification, the posterior mean of the log dividend yield ranges from -3.25 to -3.40. The sample mean is -3.54. It follows that the shocks v_t over the sample period must be negative on average. Because of the negative correlation between shocks to the dividend price ratio and to expected returns, the shocks u_t must be positive on average. Therefore the posterior mean lies below the sample mean.

Continuing with the exact likelihood case, the posterior mean of μ_x is highest (and hence furthest from the sample mean) in the no-predictability model, and becomes lower as the prior becomes less dogmatic. Excess returns follow this pattern in reverse, namely they are lowest (and furthest from the sample mean) for the no-predictability model and highest for the predictability model with the least dogmatic prior. This effect may arise from the persistence ρ. The more dogmatic the prior, the closer the posterior mean of the persistence is to one. The more persistent the process, the more likely the positive shocks are to accumulate, and the more the sample mean is likely to deviate from the true posterior mean.

The results are very different when the conditional likelihood is used, as shown in Panel B. For the no-predictability model, $\mu_r = \alpha$ is equal to the sample mean. However, as long as there is some predictability, estimation of μ_r depends on μ_x, which is unstable due to the presence of $1-\rho$ in the denominator. It is striking that, in contrast to our main specification, the conditional likelihood specification has great difficultly in pinning down the mean of expected excess stock returns.

3.4 The posterior distribution

We now examine the posterior probability that excess returns are predictable. For convenience, we present results for our main specification that uses the exact likelihood. As a first step, we examine the posterior distribution for the R^2.

The posterior distribution of the R^2

Figure 2 shows two plots on the prior and posterior distribution of the R^2 with priors $P(R^2 > 1\% \,|\, H_1) = 0.50$ and $q = 0.5$. Panel A plots $P(R^2 > k)$ as a function of k for both the prior and the posterior; this corresponds to 1 minus the cumulative density function of the R^2.[7] The plot for the $P(R^2 > k)$ demonstrates a rightward shift for the posterior for values of k below (roughly) 2%.

The strength of the predictability can be seen in that while the prior implies $P(R^2 > 1\%) = 0.25$, the posterior implies $P(R^2 > 1\%)$ close to 0.50. Thus, after observing the data, an investor revises his beliefs on the strength of predictability substantially upward. Panel B plots the probability density function of the R^2. The prior places the highest density on low values of the R^2. The posterior however places high density in the region around 2% and has lower density than the prior for R^2 values close to zero. The evidence in favor of predictability, with a moderate R^2, is sufficient to overcome the investor's initial skepticism.

The posterior probability of return predictability

Figure 2 shows the posterior R^2 for a given set of prior beliefs. Table 2 shows how various statistics on the posterior distribution vary as the prior distribution changes. Table 2 presents the posterior probabilities of predictability as a function of the investor's prior about the existence of predictability, q, and the prior belief on the strength of predictability. The posterior probability is increasing in q and hump-shaped in the strength of the prior, reflecting the fact that the Bayes factors are hump-shaped in the strength of the prior. The results demonstrate that investors with moderate beliefs on both the existence and strength of predictability revise their beliefs on the existence on predictability sharply upward. For example, an investor with $q = 0.5$ and $P(R^2 > .01|H_1) = 0.50$ conclude that the posterior likelihood of predictability equals 0.86. This result is robust to a wide range of choices for $P(R^2 > .01|H_1)$. As the table shows, $P(R^2 > .01|H_1) = 0.25$ implies a posterior probability of 0.87. The posterior probability falls off dramatically as $P(R^2 > .01|H_1)$ approaches one; for these very diffuse priors (which imply what might be considered an economically unreasonable amount of predictability), the Bayes factors are close to zero. Table 2 also

[7] This figures shows the unconditional posterior probability that the R^2 exceeds k; that is, it does not condition on the existence of predictability.

shows reasonably high means of the β and the R^2, except for the diffuse prior.

The above analysis evaluates the statistical evidence on predictability. The Bayesian approach also enables us to study the economic gains from market timing. In particular, we can evaluate the certainty equivalent loss from failing to time the market under different priors on the existence and strength of predictability.

Certainty equivalent returns

We now measure the economic significance of the predictability evidence using certainty equivalent returns. We assume an investor who maximizes

$$E\left[\left.\frac{W_{T+1}^{1-\gamma}}{1-\gamma}\,\right|\, D\right]$$

for $\gamma = 5$, where $W_{T+1} = W_T(w\exp\{r_{T+1} + r_{f,T}\} + (1-w)\exp\{r_{f,T}\})$, and w is the weight on the risky asset. The expectation is taken with respect to the predictive distribution

$$p(r_{T+1}\,|\,D) = \bar{q}p(r_{T+1}\,|\,D, H_1) + (1-\bar{q})p(r_{T+1}\,|\,D, H_0),$$

where

$$p(r_{T+1}\,|\,D, H_i) \quad = \quad \int p(r_{T+1}\,|\,x_T, b_i, \Sigma, H_i)p(b_i, \Sigma\,|\,D, H_i)\,db_i\,d\Sigma$$

for $i = 0, 1$.

A draw r_{T+1} from the distribution $p(r_{T+1}\,|\,x_T, b_1, \Sigma)$ is given by (1) with probability $\bar{q}$ and (2) with probability $1-\bar{q}$. The posterior distribution of the parameters is described in Section 2.4.

For any portfolio weight w, we can compute the certainty equivalent return as solving

$$\frac{\exp\{(1-\gamma)\mathrm{CER}\}}{1-\gamma} = E\left[\left.\frac{(w\exp\{r_{T+1}+r_{f,T}\} + (1-w)\exp\{r_{f,T}\})^{1-\gamma}}{1-\gamma}\,\right|\, D\right]. \qquad (28)$$

Following Kandel and Stambaugh (1996), we measure utility loss as the difference between certainty equivalent returns from following the optimal strategy and from following a sub-optimal strategy. We define the sub-optimal strategy as the strategy that the investor would follow if he believes that there is no predictability. Note, however, that the expectation in (28) is computed with respect to the same distribution for both the optimal and sub-optimal strategy.

Panel D of Table 2 shows the difference in certainty equivalent returns as described above. These differences are averaged over the posterior distribution for x to create a difference that is not conditional on any specific value. The data indicate economically meaningful economic losses from failing to time the market. Panel D shows that, for example, an investor with a prior on β such that $P(R^2 > .01|H_1) = 0.50$ and a 50% prior belief in the existence of return predictability would suffer a certainty equivalent loss of 1.72% (in annual terms) from failing to time the market. Higher values of q imply greater certainty equivalent losses.

3.5 Evolution of the posterior distribution over time

We next describe the evolution of the posterior distribution over time. This distribution exhibits surprising behavior over the 2000-2005 period. This behavior is a direct result of the stochastic properties of the predictor variable x_t.

Starting in 1972, we compute the posterior distribution conditional on having observed data up to and including that year. We start in 1972 because this allows for twenty years of data for the first observation. The posterior is computed by simulating 500,000 draws and dropping the first 100,000. To save on computation time, we update the posterior every year. For reference, Figure 3 shows the time-series of the log dividend-price ratio. As we will see, much of the behavior of the posterior distribution can be understood based on the time series of this ratio.

Figure 4 shows the posterior probability of predictability ($\bar{q}$) in Panel A (assuming a prior probability of 0.5). The solid line corresponds to our benchmark specification. This line is above 90% for most of the sample (it is actually at its lowest value at the end of the sample). In the 2000-2005 period, the probability is not distinguishable from one. This is surprising: intuition would suggest that the period in which the dividend-price ratio was rising far above its long-run mean (and during which returns kept being high despite these high levels) would correspond to an exceptionally low posterior probability of predictability, not a high one. Indeed, it is surprising that data could ever lead the investor to a nearly 100% certainty about the predictability model.

Panel B, which shows the Bayes factors, gives another perspective on this result. Between 2000 and 2005, the Bayes factor in favor of predictability rises to values that dwarf any others during the sample. The posterior probability takes these Bayes factors and maps them to the [0, 1] interval, so values as high as those shown in the figure are translated to posterior probabilities extremely close to 1. Why is it that the Bayes factors rise so high?

An answer is suggested by the time series behavior of β and ρ, shown in Figure 5. The solid lines show the posterior distributions of β and ρ.[8] The dashed line shows OLS estimates. The posterior for β lies below the OLS estimate for most of the period, while the posterior for ρ lies above the OLS estimator for most of the period. An exception occurs in 2001, when the positions reverse. The posterior for β lies above the OLS estimate and the posterior for ρ lies below it. Note that the OLS estimate of β is biased upwards and the OLS estimate of ρ is biased downwards, so this switch is especially surprising.

The fact that the posterior ρ rises to meet the OLS ρ, and even exceeds it, indicates that the model interprets the rise of the dividend-price ratio as occurring because of an unusual sequence of positive shocks v_t. Namely, positive shocks are more likely to occur after positive shocks during this period. This implies that negative shocks to u_t are also more likely to follow positive shocks v_t than they should, so OLS will in fact underestimate the true β (or it will overestimate the true β by less than usual).

This result is similar in spirit to that found in the frequentist analysis of Lewellen (2004) and Campbell and Yogo (2006) (see also the discussion in the survey article Campbell (2008)).

[8]For the argument below, it makes the most sense, strictly speaking, to examine the posterior distribution of β conditional on the predictability model. However, because the posterior probability of this model is so close to one, this conditional posterior β is nearly indistinguishable from the unconditional posterior β. The same is true for posterior ρ. Therefore, for simplicity, we focus on the unconditional posterior.

It is also an example of how information about shocks that are correlated with errors from a forecasting model can help improve forecasts, as in Faust and Wright (2011). Figure 4 shows that the consequences of this result for model selection are quite large. This is because the no-predictability model implies, of course, that β is zero. However, given that OLS finds a positive β, the no-predictability model implies that positive shocks to the dividend-price ratio were follows by positive shocks to returns. This is extremely unlikely, given the time series. Thus the evidence comes to strongly favor the predictability model.[9]

This chain of inference requires knowledge of the behavior of shocks to the predictor variable. The fixed-regressor approach described above eliminates such knowledge and leads to completely different inference over this time period. To illustrate this result, we compute the posterior probability of return predictability (by the dividend-price ratio), and the Bayes factor, using the prior-likelihood combination of Cremers (2002).[10] Very similar results are obtained for the methods in Avramov (2002). The probability of predictability computed using the fixed-regressor approach indeed shows a decline over the 2000-2005 period, stemming from the decline in the OLS estimate of β.

Finally, to demonstrate the impact of these prior beliefs on portfolio choice, we compute the optimal weight in the risky asset for the power utility investor described in Section 3.4. We consider four specifications of the prior. Panel A assumes that $P(R^2 > 0.01|H_1) = 0.05$, so that predictability, if it exists, is weak. Posteriors are computed for $q = 0.01$ and $q = 0.99$. Panel B repeats this exercise for $P(R^2 > 0.01|H_1) = 0.50$. Both panels show that the investor with a strong prior that returns are not predictable ($q = .01$) engages in notably less market timing that an investor who is convinced that there is some predictability ($q = .99$). Comparing Panel A with Panel B shows that the investor who believes that predictability is very weak (assuming it exists) times the market much less than an investor with a somewhat

[9] We also performed this analysis using the conditional rather than the exact likelihood. Because the results are qualitatively the same, we do not present them here.

[10] We use the priors corresponding to Cremers "confident" view. Varying the choice of prior parameters does not impact the qualitative result that we present.

flatter (though still highly informative) prior. Consider the investor with $q = 0.01$ and $P(R^2 > 0.01|H_1) = 0.05$ (the solid line in Panel A). This investor has a prior believe that there is only a 1% chance that returns are predictable. Even if returns are predictable, he believes the predictability must be very weak: there is only a 5% chance that the R^2 from a quarterly regression exceeds 1%. However, even this investor finds overwhelming evidence for stock return predictability in the 2002–2005 period.

4 Conclusion

This study has taken a Bayesian approach to the question of whether the equity premium varies over time. We considered investors who face uncertainty both over whether predictability exists, and over the strength of predictability if it does exist. We found substantial evidence in favor of predictability when the dividend-price ratio is used to predict returns. Moreover, we found large certainty equivalent losses from failing to time the market, even for investors who have strong prior beliefs in a constant equity premium.

When we examined the time series of the investor's beliefs, we found that the belief in return predictability goes to one in the 2000-2005 period, even for investors who place a very low prior on return predictability. This surprising result is a consequence of correctly modeling the regressor as stochastic rather than fixed. We also found that our posterior mean return is notably different from the sample average, a result that again stems from taking the stochastic nature of the regressor into account. These results demonstrate that the way that the regressor is modeled is very important. We hope to examine this issue further in future work.

Appendix

A Jeffreys prior under H_0

Our derivation for the limiting Jeffreys prior on b_0, Σ generalizes that of Stambaugh (1999). Zellner (1996, pp. 216-220) derives a limiting Jeffreys prior by applying (A.1) to the likelihood (17) and retaining terms of the highest order in T. Stambaugh shows that Zellner's approach is equivalent to applying (A.1) to the conditional likelihood (16), and taking the expectation in (A.1) assuming that x_0 is multivariate normal with mean (6) and variance (7). We adopt this approach.

Given a set of parameters μ, data D, and a log-likelihood $l(\mu; D)$, the limiting Jeffreys prior satisfies

$$p(\mu) \propto \left| -E\left(\frac{\partial^2 l}{\partial \mu \partial \mu^\top}\right) \right|^{1/2}. \tag{A.1}$$

We derive the prior density for $p(b_0, \Sigma^{-1})$ and then transform this into the density for $p(b_0, \Sigma)$ using the Jacobian. Let

$$l_0(b_0, \Sigma; D) = \log p(D | b_0, \Sigma, H_0, x_0). \tag{A.2}$$

denote the natural log of the conditional likelihood. Let $\zeta = [\sigma^{(11)}\ \sigma^{(12)}\ \sigma^{(22)}]^\top$, where $\sigma^{(ij)}$ denotes element (i,j) of Σ^{-1}. Applying (A.1) implies

$$p(b_0, \Sigma^{-1} | H_0) \propto \left| -E \begin{bmatrix} \frac{\partial^2 l_0}{\partial b_0 \partial b_0^\top} & \frac{\partial^2 l_0}{\partial b_0 \partial \zeta^\top} \\ \frac{\partial^2 l_0}{\partial \zeta \partial b_0^\top} & \frac{\partial^2 l_0}{\partial \zeta \partial \zeta^\top} \end{bmatrix} \right|^{1/2}. \tag{A.3}$$

The form of the conditional likelihood implies that

$$l_0(b_0, \Sigma; D) = -\frac{T}{2} \log |2\pi\Sigma| - \frac{1}{2} (z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right) (z - Z_0 b_0). \tag{A.4}$$

It follows from (A.4) that

$$\frac{\partial l_0}{\partial b_0} = \frac{1}{2} Z_0^\top \left(\Sigma^{-1} \otimes I_T\right) (z - Z_0 b_0),$$

and

$$
\begin{aligned}
\frac{\partial^2 l_0}{\partial b_0 \partial b_0^\top} &= -\frac{1}{2} Z_0^\top \left(\Sigma^{-1} \otimes I_T\right) Z_0 \\
&= -\frac{1}{2} \begin{bmatrix} \iota_T^\top & 0 \\ 0 & X^\top \end{bmatrix} \left(\Sigma^{-1} \otimes I_T\right) \begin{bmatrix} \iota_T & 0 \\ 0 & X \end{bmatrix} \\
&= -\frac{1}{2} \begin{bmatrix} \sigma^{(11)} T & \sigma^{(12)} \iota^\top X \\ \sigma^{(12)} X^\top \iota & \sigma^{(22)} X^\top X \end{bmatrix}. \qquad \text{(A.5)}
\end{aligned}
$$

Taking the expectation conditional on b_0 and Σ implies

$$
E\left[\frac{\partial^2 l_0}{\partial b_0 \partial b_0^\top}\right] = -\frac{T}{2} \begin{bmatrix} \sigma^{(11)} & \sigma^{(12)} [1 \ \mu_x] \\ \sigma^{(12)} \begin{bmatrix} 1 \\ \mu_x \end{bmatrix} & \sigma^{(22)} \begin{bmatrix} 1 & \mu_x \\ \mu_x & \sigma_x^2 + \mu_x^2 \end{bmatrix} \end{bmatrix} \qquad \text{(A.6)}
$$

Using arguments in Stambaugh (1999), it can be shown that

$$
E\left[\frac{\partial^2 l_0}{\partial b_0 \partial \zeta^\top}\right] = 0.
$$

Moreover,

$$
-\left| E\left(\frac{\partial^2 l_0}{\partial \zeta \partial \zeta^\top}\right) \right| = \left| \frac{\partial^2 \log |\Sigma|}{\partial \zeta \partial \zeta^\top} \right| = |\Sigma|^3
$$

(see Box and Tiao (1973, pp. 474-475)). Therefore

$$
p(b_0, \Sigma^{-1} | H_0) \propto |\Phi|^{\frac{1}{2}} |\Sigma|^{\frac{3}{2}} \qquad \text{(A.7)}
$$

where

$$
\Phi = \begin{bmatrix} \Sigma^{-1} & \mu_x \begin{bmatrix} \sigma^{(12)} \\ \sigma^{(22)} \end{bmatrix} \\ \mu_x \left[\sigma^{(12)} \ \sigma^{(22)}\right] & \left(\sigma_x^2 + \mu_x^2\right) \sigma^{(22)} \end{bmatrix}.
$$

This matrix Φ has the same determinant as $-E\left[\frac{\partial^2 l_0}{\partial b_0 \partial b_0^\top}\right]$ because 2 columns and 2 rows have been reversed.

From the formula for the determinant of a partitioned matrix, it follows that

$$
|\Phi| = \left|\Sigma^{-1}\right| \left| \left(\sigma_x^2 + \mu_x^2\right) \sigma^{(22)} - \mu_x^2 \left[\sigma^{(12)} \ \sigma^{(22)}\right] \Sigma \begin{bmatrix} \sigma^{(12)} \\ \sigma^{(22)} \end{bmatrix} \right|.
$$

Because

$$\Sigma \begin{bmatrix} \sigma^{(12)} \\ \sigma^{(22)} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

it follows that

$$\begin{aligned} |\Phi| &= \left|\Sigma^{-1}\right| \left|\left(\sigma_x^2 + \mu_x^2\right)\sigma^{(22)} - \mu_x^2 \sigma^{(22)}\right| \\ &= |\Sigma|^{-1}\sigma_x^2 \sigma^{(22)}. \end{aligned}$$

The determinant of Σ equals

$$|\Sigma| = \sigma_u^2 \left(\sigma_v^2 - \sigma_{uv}^2 \sigma_u^{-2}\right),$$

while $\sigma^{(22)} = (\sigma_v^2 - \sigma_{uv}^2 \sigma_u^{-2})^{-1}$. Therefore,

$$|\Phi| = |\Sigma|^{-2}\sigma_u^2 \sigma_x^2.$$

Substituting into (A.7),

$$p(b_0, \Sigma^{-1}|H_0) \propto |\Sigma|^{\frac{1}{2}}\sigma_u \sigma_x.$$

The Jacobian of the transformation from Σ^{-1} to Σ is $|\Sigma|^{-3}$. Therefore,

$$p(b_0, \Sigma|H_0) = |\Sigma|^{-\frac{5}{2}}\sigma_u \sigma_x.$$

B Sampling from Posterior Distributions

This section describes how to sample from the posterior distributions. In all cases, the sampling procedure for the posteriors under H_1 and H_0 involve the Metropolis-Hastings algorithm. Below we describe the case of the exact likelihood in detail. The procedure for the conditional likelihood is similar.

B.1 Posterior distribution under H_0

Substituting (8) and (17) into (18) implies that

$$p(b_0, \Sigma|H_0, D) \propto \sigma_u |\Sigma|^{-\frac{T+5}{2}} \exp\left\{-\frac{1}{2}\sigma_x^{-2}(x_0 - \mu_x)^2 - \frac{1}{2}(z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_0 b_0)\right\}.$$

This posterior does not take the form of a standard density function because of the term in the likelihood involving x_0 (note that σ_x^2 is a nonlinear function of ρ and σ_v). However, we can sample from the posterior using the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm is implemented "block-at-a-time", by repeatedly sampling from $p(\Sigma|b_0, H_0, D)$ and from $p(b_0|\Sigma, H_0 D)$ and repeating. To calculate a proposal density for Σ, note that

$$(z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_0 b_0) = \text{tr}\left[(Y - XB_0)^\top (Y - XB_0)\Sigma^{-1}\right],$$

where

$$B_0 = \begin{bmatrix} \alpha & \theta \\ 0 & \rho \end{bmatrix}.$$

The proposal density for the conditional probability of Σ is the inverted Wishart with $T + 2$ degrees of freedom and scale factor of $(Y - XB_0)^\top (Y - XB_0)$. The target is therefore

$$p(\Sigma|b_0, H_0, D) \propto \sigma_u \exp\left\{-\frac{1}{2}(x_0 - \mu_x)^2 \sigma_x^{-2}\right\} \times \text{proposal.}$$

Let

$$V_0 \quad = \quad \left(Z_0^\top \left(\Sigma^{-1} \otimes I_T\right) Z_0\right)^{-1}$$

Let

$$\hat{b}_0 = V_0 Z_0^\top \left(\Sigma^{-1} \otimes I_T\right) z$$

It follows from completing the square that

$$(z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_0 b_0) = (b_0 - \hat{b}_0)^\top V_0^{-1}(b_0 - \hat{b}_0) + \text{ terms independent of } b_0.$$

The proposal density for b_0 is therefore multivariate normal with mean $\hat{b}_0$ and variance-covariance matrix V_0. The accept-reject algorithm of Chib and Greenberg (1995, Section 5) is used to sample from the target density, which is equal to

$$p(b_0|\Sigma, H_0, D) \propto \exp\left\{-\frac{1}{2}(x_0 - \mu_x)^2 \sigma_x^{-2}\right\} \times \text{proposal.}$$

Note that σ_u and Σ are in the constant of proportionality. Drawing successively from the conditional posteriors for Σ and b_0 produces a density that converges to the full posterior conditional on H_0.

B.2 Posterior distribution under H_1

Substituting (12) and (15) into (18) implies that

$$p(b_1, \Sigma|H_1, D) \propto \sigma_x |\Sigma|^{-\frac{T+5}{2}} \exp\left\{-\frac{1}{2}\beta^2 \left(\sigma_\eta^2 \sigma_x^{-2} \sigma_u^2\right)^{-2} - \frac{1}{2}\sigma_x^{-2}(x_0 - \mu_x)^2\right\}$$
$$\exp\left\{-\frac{1}{2}(z - Z_1 b_1)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_1 b_1)\right\}.$$

The sampling procedure is similar to that described in Appendix B.1. Details can be found in Wachter and Warusawitharana (2009). To summarize, we first draw from the posterior $p(\Sigma \,|\, b_1, H_1, D)$. The proposal density is an inverted Wishart with $T + 2$ degrees of freedom and scale factor $(Y - XB_1)^\top(Y - XB_1)$, where

$$B_1 = \begin{bmatrix} \alpha & \theta \\ \beta & \rho \end{bmatrix}.$$

We then draw from $p(\theta, \rho \,|\, \alpha, \beta, \Sigma, H_1, D)$. The proposal density is multivariate normal with mean and variance determined by the conditional normal distribution. Finally, we draw from $p(\alpha, \beta \,|\, \theta, \rho, \Sigma, H_1, D)$. In this case, the target and the proposal are the same, and are also multivariate normal.

C Computing the Bayes factor

Verdinelli and Wasserman (1995) provide an implementable formula for the inverse of the Bayes factor. In our notation, this formula can be written as

$$\mathcal{B}_{10}^{-1} = p(\beta = 0|H_1, D) E\left[\frac{p(b_0, \Sigma|H_0)}{p(\beta = 0, b_0, \Sigma|H_1)} \,\middle|\, \beta = 0, H_1, D\right]. \tag{C.1}$$

To compute $p(\beta = 0 \,|\, H_1, D)$, note that

$$p(\beta = 0 \,|\, H_1, D) = \int p(\beta = 0 \,|\, b_0, \Sigma, H_1, D) p(b_0, \Sigma \,|\, H_1, D)\, db_0\, d\Sigma. \tag{C.2}$$

As discussed in Appendix B.2, the posterior distribution of α and β conditional on the remaining parameters is normal. We can therefore compute $p(\beta = 0 \,|\, b_0, \Sigma, H_1, D)$ (including integration constants) in closed form, by using the properties of the conditional normal distribution. Consider N draws from the full posterior: $((b_1^{(1)}, \Sigma^{(1)}), \ldots, (b_1^{(N)}, \Sigma^{(N)}))$, where we can write $(b_1^{(i)}, \Sigma^{(i)})$ as $(\beta^{(i)}, b_0^{(i)}, \Sigma^{(i)})$. We use these draws to integrate out over b_0 and Σ. It follows from (C.2) that

$$p(\beta = 0|H_1, D) \approx \frac{1}{N}\sum_{i=1}^{N} p(\beta = 0|b_0^{(i)}, \Sigma^{(i)}, H_1, D).$$

where the approximation is accurate for large N.

To compute the second term in (C.1), we observe that

$$\frac{p(b_0, \Sigma \,|\, H_0)}{p(\beta = 0, b_0, \Sigma \,|\, H_1)} = \frac{p(b_0, \Sigma \,|\, H_0)}{p(\beta = 0|b_0, \Sigma, H_1)p(b_0, \Sigma \,|\, H_1)} = \sqrt{2\pi}\sigma_\beta,$$

because $p(b_0, \Sigma \,|\, H_0) = p(b_0, \Sigma \,|\, H_1)$. Note that $\sigma_\beta = \sigma_\eta \sigma_x^{-1} \sigma_u$. We require the expectation taken with respect to the posterior distribution conditional on the existence of predictability and the realization $\beta = 0$. To calculate this expectation, we draw $((b_0^{(1)}, \Sigma^{(1)}), \ldots, (b_0^{(N)}, \Sigma^{(N)}))$ from $p(b_0, \Sigma \,|\, \beta = 0, H_1, D)$. This involves modifying the procedure for drawing from the posterior for b_1, Σ given H_1 (see Appendix B.2). We sample from $p(\Sigma \,|\, \alpha, \beta = 0, \theta, \rho, H_1, D)$, then from $p(\rho, \theta \,|\, \alpha, \beta = 0, \Sigma, H_1, D)$ and finally from $p(\alpha \,|\, \beta = 0, \Sigma, \theta, \rho, H_1, D)$, and repeat until the desired number of draws are obtained. All steps except the last are identical to those described in Appendix B.2 (the value of β is identically zero rather than the value from the previous draw). For the last step we derive $p(\alpha \,|\, \beta = 0, \Sigma, \theta, \rho, H_1, D)$ from the joint distribution $p(\alpha, \beta \,|\, \Sigma, \theta, \rho, H_1, D)$, making use of the properties of the conditional normal distribution.

Given these draws from the posterior distribution, the second term equals

$$E\left[\frac{p(b_0, \Sigma|H_0)}{p(\beta = 0, b_0, \Sigma|H_1)} \,\middle|\, \beta = 0, H_1, D\right] \approx \frac{1}{N}\sum_{i=1}^{N} \sqrt{2\pi}\sigma_\eta (\sigma_x^{(i)})^{-1} \sigma_u^{(i)}, \tag{C.3}$$

where this approximation is accurate for N large.

References

Ando, Tomohiro, and Arnold Zellner, 2010, Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques, *Bayesian Analysis* 5, 65–96.

Avramov, Doron, 2002, Stock return predictability and model uncertainty, *Journal of Financial Economics* 64, 423–458.

Barberis, Nicholas, 2000, Investing for the long run when returns are predictable, *Journal of Finance* 55, 225–264.

Bartlett, M.S., 1957, Comment on 'A Statistical Paradox' by D. V. Lindley, *Biometrika* 44, 533–534.

Box, George E.P., and George C. Tiao, 1973, *Bayesian Inference in Statistical Analysis*. (Addison-Wesley Pub. Co. Reading, MA).

Brandt, Michael W., Amit Goyal, Pedro Santa-Clara, and Jonathan R. Stroud, 2005, A simulation approach to dynamics portfolio choice with an application to learning about return predictability, *Review of Financial Studies* 18, 831–873.

Campbell, John Y., 2008, Viewpoint: Estimating the equity premium, *Canadian Journal of Economics* 41, 1–21.

Campbell, John Y., and Motohiro Yogo, 2006, Efficient tests of stock return predictability, *Journal of Financial Economics* 81, 27–60.

Chen, Zengjing, and Larry Epstein, 2002, Ambiguity, risk and asset returns in continuous time, *Econometrica* 70, 1403–1443.

Chib, Siddhartha, and Edward Greenberg, 1995, Understanding the Metropolis-Hastings algorithm, *American Statistician* 49, 327–335.

Chipman, Hugh, Edward I. George, and Robert E. McCulloch, 2001, The practical implementation of Bayesian model selection, in P. Lahiri, eds.: *Model Selection* (IMS Lecture Notes, Bethesda, MA).

Cremers, K.J. Martjin, 2002, Stock return predictability: A Bayesian model selection perspective, *Review of Financial Studies* 15, 1223–1249.

Dickey, James M., 1971, The weighted likelihood ratio, linear hypotheses on normal location paramaters, *The Annals of Mathematical Statistics* 42, 204–223.

Faust, Jon, and Jonathan H. Wright, 2011, Efficient Prediction of Excess Returns, *Review of Economics and Statistics* 93, 647–659.

Fernandez, Carmen, Eduardo Ley, and Mark F. J. Steel, 2001, Benchmark priors for Bayesian model averaging, *Journal of Econometrics* 100, 381–427.

Hamilton, J. D., 1994, *Time Series Analysis.* (Oxford University Press Princeton, NJ).

Jeffreys, Harold, 1961, *Theory of Probability.* (Oxford University Press Clarenden).

Johannes, Michael, and Nicholas Polson, 2006, MCMC methods for financial econometrics, in Yacine Ait-Sahalia, and Lars Hansen, eds.: *Handbook of Financial Econometrics* (Elsevier, North-Holland).

Johannes, Michael, Nicholas Polson, and Jonathan R. Stroud, 2002, Sequential optimal portfolio performance: Market and volatility timing, Working paper, Columbia University, University of Chicago, and University of Pennsylvania.

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

Kass, R., and A. E. Raftery, 1995, Bayes factors, *Journal of the American Statistical Association* 90, 773–795.

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

Mehra, Rajnish, and Edward Prescott, 1985, The equity premium puzzle, *Journal of Monetary Economics* 15, 145–161.

Pastor, Lubos, and Robert F. Stambaugh, 2009, Predictive systems: Living with imperfect predictors, *Journal of Finance* 64, 1583 – 1628.

Poirier, Dale J., 1978, The effect of the first observation in regression models with first-order autoregressive disturbances, *Journal of the Royal Statistical Society, Series C, Applied Statistics* 27, 67–68.

Shanken, Jay A., and Ane Tamayo, 2011, Payout yield, risk and mispricing, a Bayesian analysis, forthcoming, *Journal of Financial Economics.*

Skoulakis, Georgios, 2007, Dynamic portfolio choice with Bayesian learning, Working paper, University of Maryland.

Stambaugh, Robert F., 1999, Predictive regressions, *Journal of Financial Economics* 54, 375–421.

Stock, James H., and Mark W. Watson, 2011, Generalized shrinkage methods for forecasting using many predictors, Working paper, Harvard University and Princeton University.

Verdinelli, Isabella, and Larry Wasserman, 1995, Computing Bayes factors using a generalization of the Savage-Dickey density ratio, *Journal of the American Statistical Association* 90, 614–618.

Wachter, Jessica A., 2010, Asset Allocation, *Annual Review of Financial Economics* 2, 175–206.

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

Wright, Jonathan H., 2008, Bayesian Model Averaging and exchange rate forecasts, *Journal of Econometrics* 146, 329 – 341.

Zellner, Arnold, 1996, *An introduction to Bayesian inference in econometrics.* (John Wiley and Sons, Inc. New York, NY).

Table 1: Bayes factors and conditional posterior means

		Posterior Means			
$P(R^2 > 0.01 \mid H_1)$	Bayes factor	β	ρ	μ_r	μ_x
Panel A: Exact likelihood					
0	Undefined	0	0.997	3.45	-3.25
0.05	4.13	1.07	0.989	3.77	-3.35
0.25	6.48	1.65	0.985	3.85	-3.38
0.50	6.13	1.91	0.983	3.88	-3.39
0.99	0.01	2.06	0.982	3.90	-3.40
Panel B: Conditional likelihood					
0	Undefined	0	0.998	4.48	-6.83
0.05	2.00	0.74	0.993	3.70	-5.28
0.25	2.71	1.36	0.988	3.39	-4.79
0.50	2.56	1.66	0.985	3.11	-4.78
0.99	0.01	1.80	0.984	2.15	-5.03
Panel C: Ordinary least squares					
		2.97	0.973	4.49	-3.54

Notes: The Bayes factor equals the probability of the data D given the predictability model H_1 divided by the probability of the data given the no-predictability model H_0: $p(D|H_1)/p(D|H_0)$. Bayes factors are reported for various priors of the strength of predictability under H_1, indexed by $P(R^2 > 0.01|H_1)$ (namely, the prior probability that the population R^2 exceeds 0.01, assuming H_1). Posterior means are conditional on H_1 and are computed for the predictability coefficient β, the persistence of the dividend-price ratio ρ, the mean of the continuously compounded excess return μ_r, and the mean of the predictor variable μ_x. In Panel C, μ_r and μ_x equal the sample means. Data are quarterly from 7/1/1952 to 3/31/2009.

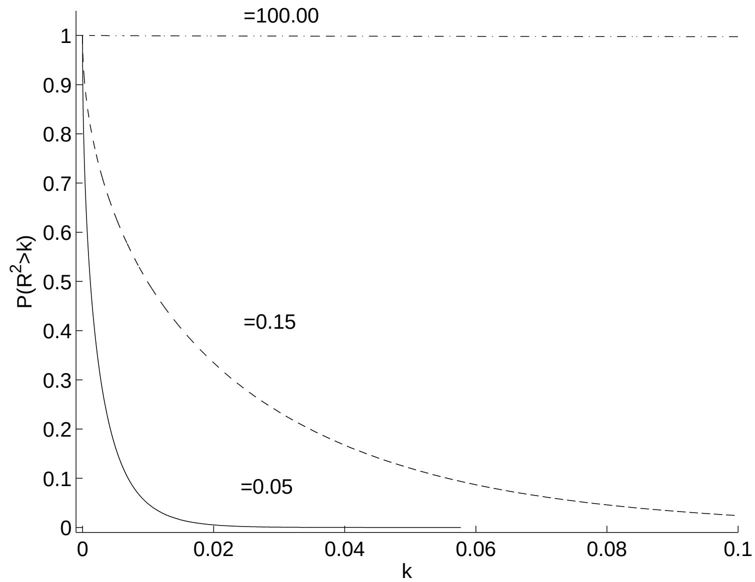
Table 2: Posterior probability of predictability, unconditional posterior means of β and R^2, and certainty equivalent returns.

$P(R^2 > 0.01 \mid H_1)$	Prior probability of return predictability q			
	0.20	0.50	0.80	0.99
Panel A: Posterior probability of predictability $\bar{q}$				
0.05	0.51	0.80	0.94	1.00
0.25	0.62	0.87	0.96	1.00
0.50	0.61	0.86	0.96	1.00
0.99	0.00	0.01	0.05	0.54
Panel B: Posterior mean of predictive coefficient β				
0.05	0.55	0.86	1.01	1.07
0.25	1.02	1.43	1.59	1.65
0.50	1.16	1.64	1.84	1.91
0.99	0.01	0.02	0.09	1.12
Panel C: Posterior mean of R^2 (in percentages)				
0.05	0.30	0.48	0.56	0.59
0.25	0.59	0.83	0.92	0.95
0.50	0.68	0.97	1.08	1.12
0.99	0.00	0.01	0.06	0.68
Panel D: Difference in CER between optimal and no-predictability strategies				
0.05	0.38	0.84	1.10	1.20
0.25	0.85	1.45	1.71	1.81
0.50	1.00	1.72	2.03	2.15
0.99	0.00	0.00	0.02	1.67

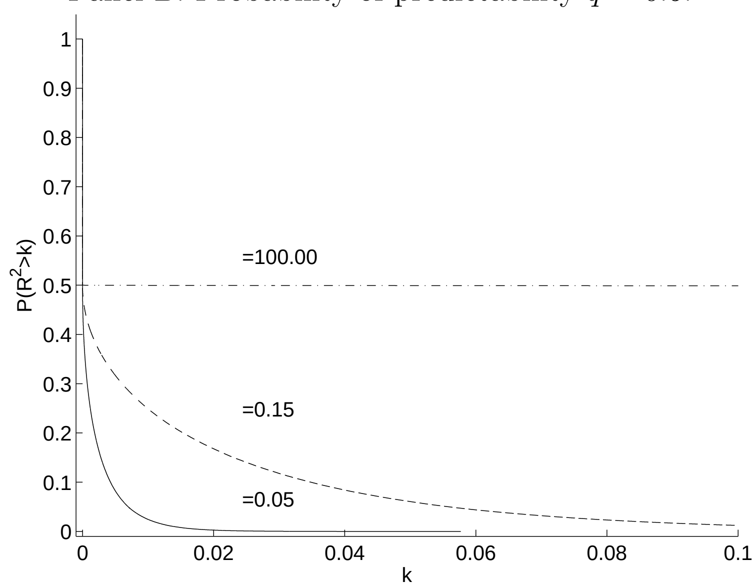
Notes: The table reports statistics of the posterior distribution, averaging over the models H_1 and H_0. The parameter q denotes the prior probability of H_1. Statistics are reported for various priors of the strength of predictability under H_1, indexed by $P(R^2 > 0.01|H_1)$ (namely, the prior probability that the population R^2 exceeds 0.01, assuming H_1). CER stands for certainty equivalent return and is annualized by multiplying by four. Data are quarterly from 7/1/1952 to 3/31/2009.

Figure 1: Prior Distribution of the R^2

Panel A: Probability of predictability $q = 1$.

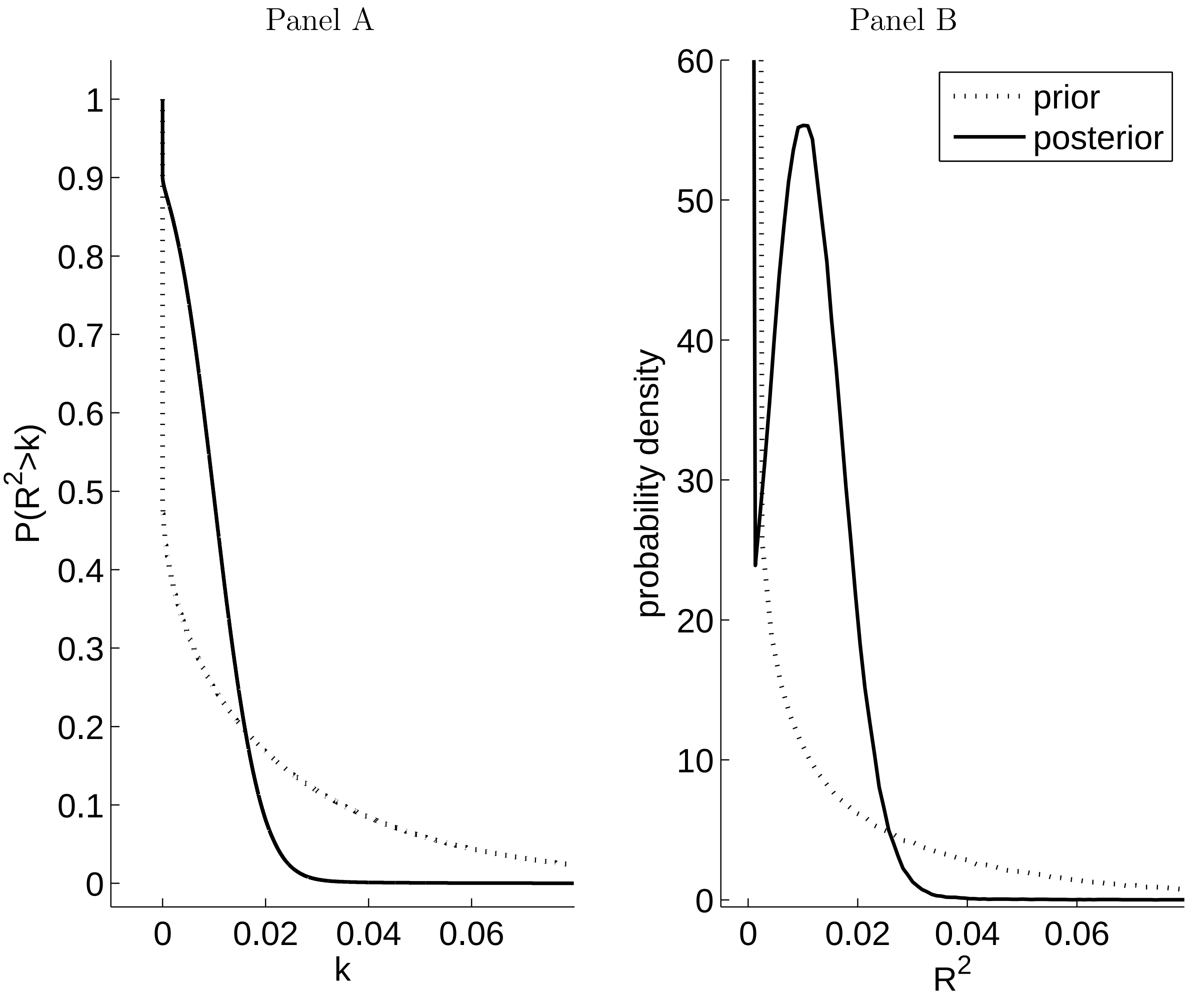


Panel B: Probability of predictability $q = 0.5$.



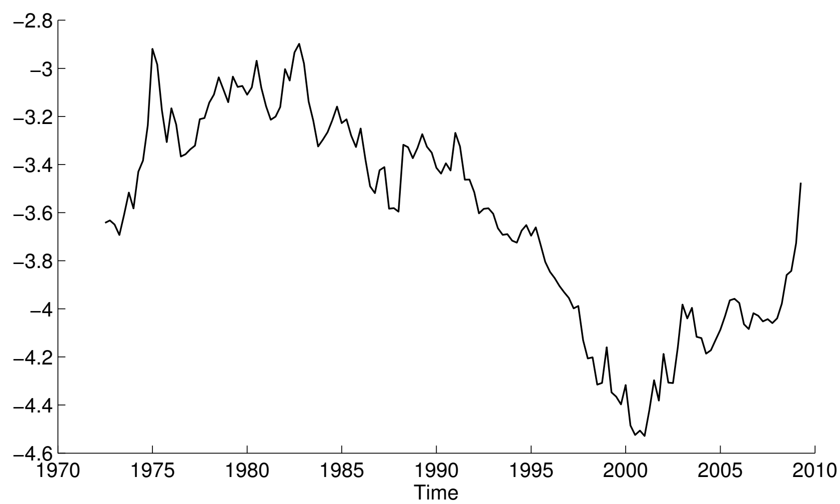
Notes: The figures plot the prior probability that the R^2 will be greater than some value k for different values of k. This equals 1 minus the cumulative density function for the distribution on the R^2. Panel A reports the values conditional on predictability ($q = 1$) and panel B plots the values for a prior value of $q = 0.5$. σ_η parameterizes the prior variance of β with $\sigma_\beta = \sigma_\eta \sigma_x^{-1} \sigma_u$.

Figure 2: Posterior Distribution of the R^2



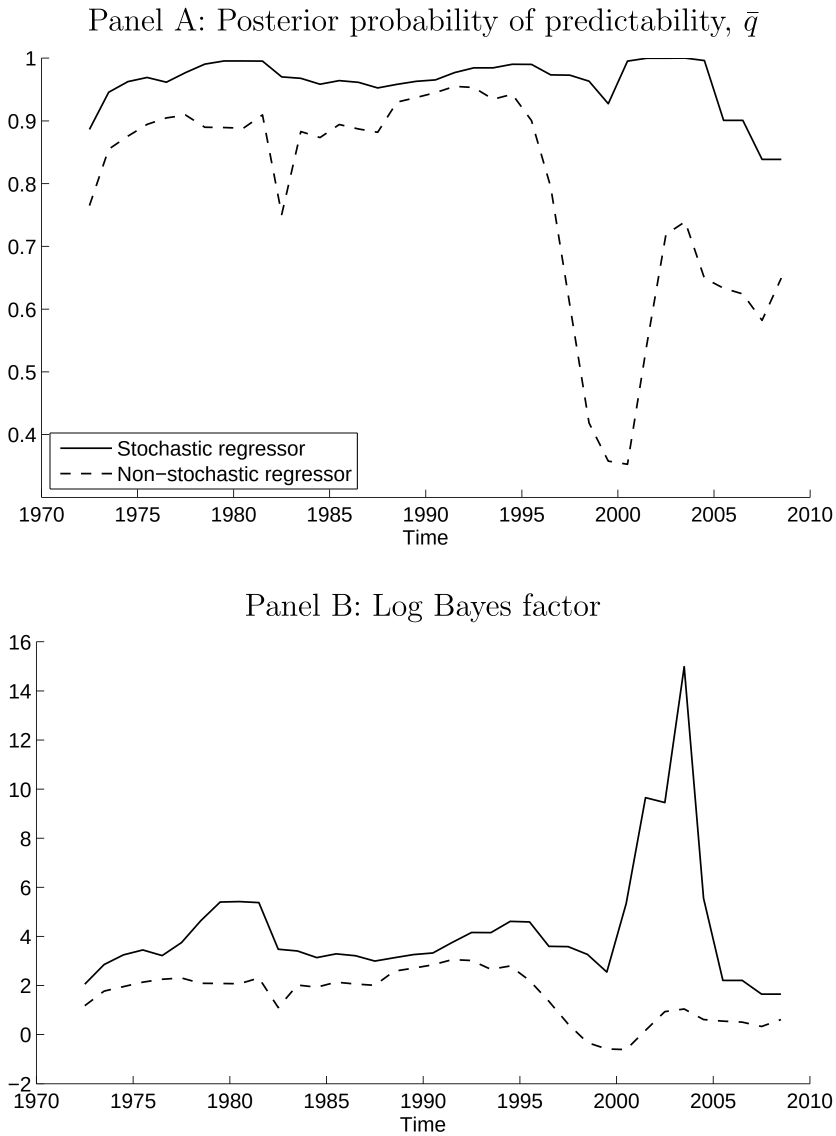
Notes: Panel A plots the probability that the R^2 from a predictive regression of excess stock returns on the payout yield will be greater than some value k for different values of k. This equals 1 minus the cumulative density function for the distribution on the R^2. Panel B plots the probability density function of the R^2 for the same regression. The dashed line signifies the prior and the solid line signifies the posterior distribution for the R^2. The likelihood function for these plots is the full Bayes exact likelihood with $P(R^2 > 0.01|H_1) = 0.50$ and $q = 0.5$. Data are quarterly from 7/1/1952 to 3/31/2009.

Figure 3: The log dividend-price ratio



Notes: The quarterly observations on the log of the dividend-price ratio, computed by dividing the dividend payout over the previous 12 months by the current price. Prices and dividends are for the CRSP value-weighted index.

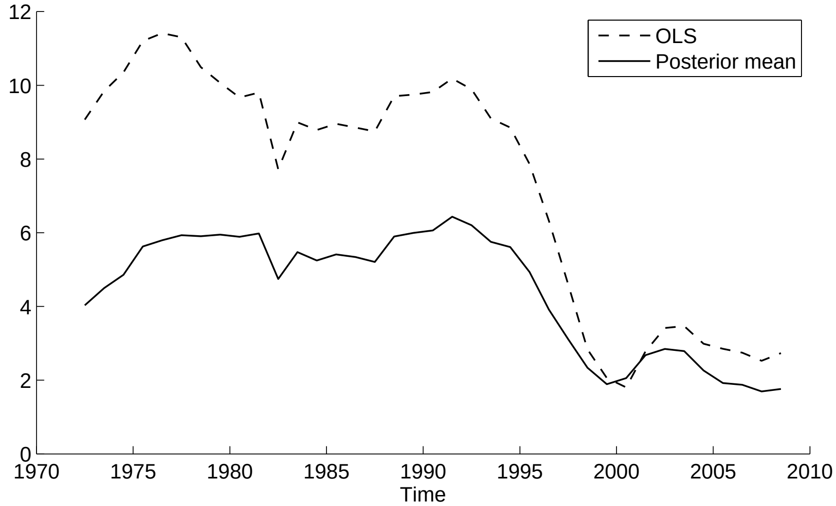
Figure 4: The Bayes factor and posterior probability of return predictability



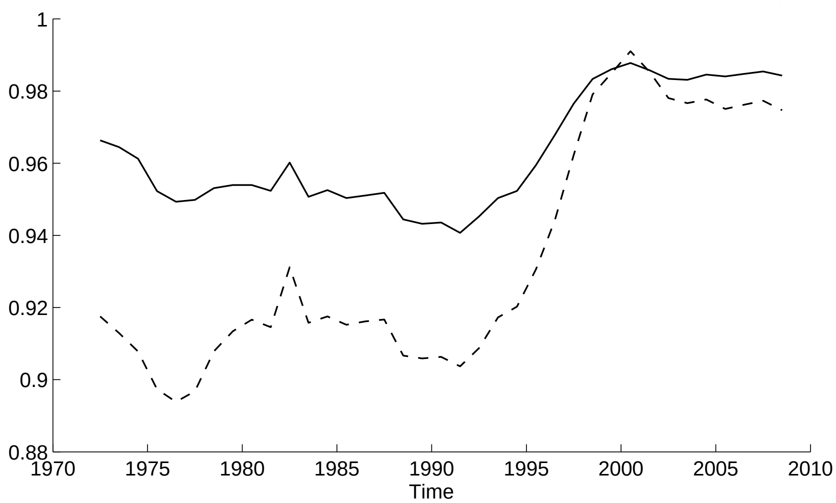
Notes: Panel A shows the posterior probability of H_1 (the predictability model), assuming a prior probability of 0.5. Panel B shows the Bayes factor, equal to the probability of the data given the predictability model H_1 divided by the probability of the data given the no-predictability model H_0. Both panels assume $P(R^2 > 0.01|H_1)$ (namely, the prior probability that the population R^2 exceeds 0.01, given H_1) is equal to 0.5. The Bayes factor and the posterior probability are computed using quarterly data beginning in 7/1/1952 and ending at the time shown on the x-axis. The solid line shows results for the main specification; the dotted line shows results for a model assuming a non-stochastic regressor.

Figure 5: Posterior means of β and ρ over time.

Panel A: Posterior mean of predictive coefficient β

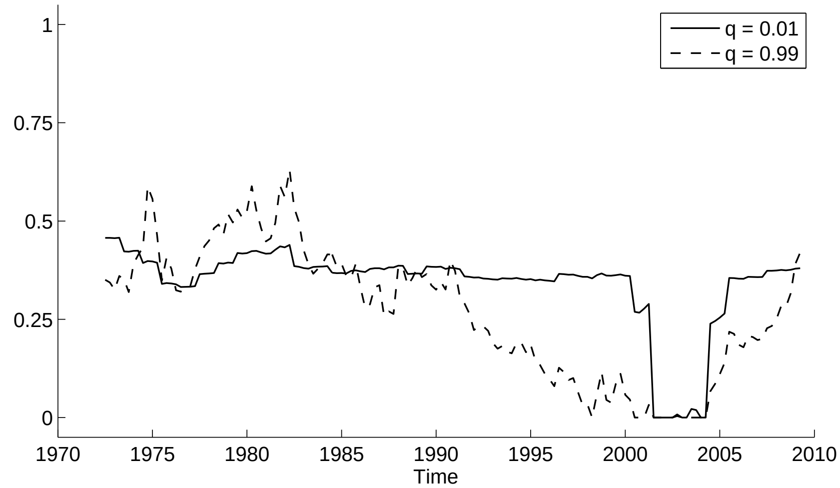


Panel B: Posterior mean of autoregressive coefficient ρ

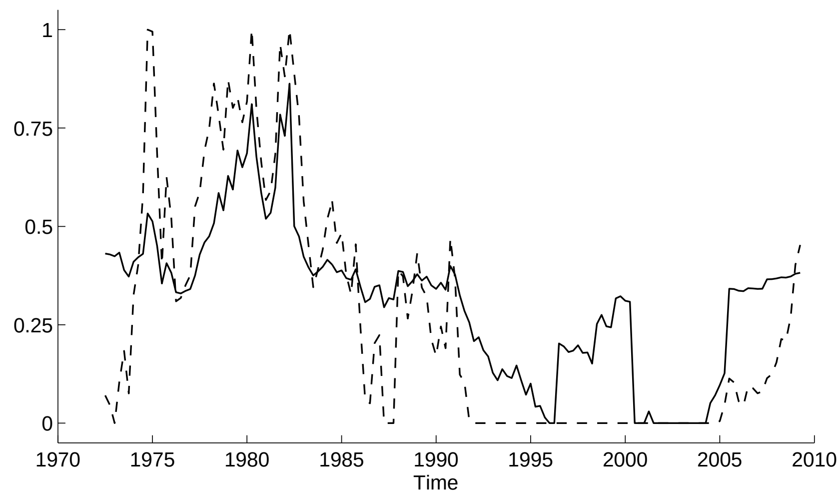


Notes: Panels show means of posterior distributions and ordinary least squares estimates. The posterior distributions are computed assuming q (the priori probability that returns are predictable) equal to 0.50, and assuming $P(R^2 > 0.01|H_1)$ (the prior probability that the population R^2 exceeds 0.01, given H_1) also equal to 0.5. The posterior distributions and OLS estimates are computing using data beginning in 7/1/1952 and ending at the time shown on the x-axis.

Panel A: Portfolio weights with $P(R^2 > 0.01|H_1) = 0.05$



Panel B: Portfolio weights with $P(R^2 > 0.01|H_1) = 0.50$



Notes: Panels show the time series of weights in the risky asset assuming prior distribution such that $P(R^2 > 0.01|H_1) = 0.05$ (Panel A) and such that $P(R^2 > 0.01|H_1) = 0.50$ (Panel B). The solid lines assume the prior probability of predictability q equals 0.01; the dashed lines assume $q = 0.99$. The investor has power utility with risk aversion equal to 5. The posterior distributions are computing using data beginning in 7/1/1952 and ending at the time shown on the x-axis.