Maximum likelihood estimation of the equity premium *

Efstathios Avdis
University of Alberta

Jessica A. Wachter
University of Pennsylvania
and NBER

December 30, 2016

Abstract

The equity premium, namely the expected return on the aggregate stock market less the government bill rate, is of central importance to the portfolio allocation of individuals, to the investment decisions of firms, and to model calibration and testing. This quantity is usually estimated from the sample average excess return. We propose an alternative estimator, based on maximum likelihood, that takes into account information contained in dividends and prices. Applied to the postwar sample, our method leads to an economically significant reduction from 6.4% to 5.1%. Simulation results show that our method produces more reliable estimates under a wide range of specifications.

*Avdis: avdis@ualberta.ca; Wachter: jwachter@wharton.upenn.edu. We are grateful to Kenneth Ahern, John Campbell, John Cochrane, Frank Diebold, Greg Duffee, Ian Dew-Becker, Adlai Fisher, Robert Hall, Soohun Kim, Alex Maynard, Ilaria Piatti, Jonathan Wright, Motohiro Yogo and seminar participants at the University of Alberta, the University of Rochester, the Wharton School, the NBER Forecasting & Empirical Methods Workshop, the SFS Cavalcade, the SoFiE Conference, the Northern Finance Association meetings and the EFA Conference for helpful comments. We thank Marco Grotteria for excellent research assistance.

1. Introduction

The equity premium, namely the expected return on equities less the risk-free rate, is an important economic quantity for many reasons. It is an input into the decision process of individual investors as they determine their asset allocation between stocks and bonds. It is also a part of cost-of-capital calculations and thus investment decisions by firms. Finally, financial economists use it to calibrate and to test, both formally and informally, models of asset pricing and of the macroeconomy.[1]

The equity premium is almost always estimated by taking the sample mean of stock returns and subtracting a measure of the riskfree rate such as the average Treasury Bill return. As is well known (Merton, 1980), it is difficult to estimate the mean of a stochastic process. A tighter estimate of the sample average cannot be obtained by sampling more finely, but rather only by extending the data series backward in time, with the disadvantage that the data are potentially less relevant to the present day.

Given the importance of the equity premium, and the noise in the sample average of stock returns, it is not surprising that a substantial literature has grown up around estimating this quantity using other methods. One idea is to use the information in dividends, given that, in the long run, prices are determined by the present value of future dividends. Studies that implement this idea in various ways include Blanchard (1993), Constantinides (2002), Donaldson et al. (2010), Fama and French (2002), and Ibbotson and Chen (2003). However, in each case it is not clear why the method in question would deliver an estimate that is superior to the sample mean.

In this paper, we propose a method of estimating the equity premium that incorporates additional information contained in the time series of prices and

[1]See, for example, the classic paper of Mehra and Prescott (1985), and surveys such as Kocherlakota (1996), Campbell (2003), DeLong and Magin (2009), and Siegel (2005).

dividends in a simple and econometrically-motivated way. As in the previous literature, our work is based on the long-run relation between prices, returns and dividends. However, our implementation is quite different, and grows directly out of maximum likelihood estimation of autoregressive processes. First, we show that our method yields an economically significant difference in the estimation of the equity premium. Taking the sample average of monthly log returns and subtracting the monthly log return on the Treasury bill over the postwar period implies a monthly equity premium of 0.43%. Our maximum likelihood approach implies an equity premium of 0.32%. Translated to level returns per annum, our method implies an equity premium of 5.06%, as compared with the sample average of 6.37%.

Second, we show that our method is a more reliable way to estimate risk premia. Because it is based on maximum likelihood, our method will be efficient in large samples. We demonstrate efficiency in small samples by running Monte Carlo experiments under a wide variety of assumptions on the data generating process, allowing for significant mis-specification. We find that the standard errors are about half as large using our method as using the sample average. We also compute the root-mean-squared error and find that it is smaller for our estimate as compared with the sample mean. These results strongly suggest that the answer given by our method is closer to the true equity premium as compared with the average return.

Finally, we are able to derive analytical expressions for our estimator that give intuition for our results. Maximum likelihood allows additional information to be extracted from the time series of the dividend-price ratio. This additional information implies that shocks to the dividend-price ratio have on average been negative. In contrast, ordinary least squares (OLS) implies that the shocks are zero on average by definition. Because shocks to the dividend-price ratio are negatively correlated with shocks to returns, our results imply that shocks to returns must have been positive over the time period. That

is, the historical time series of returns is unusually high; a lower value of the equity premium is closer to the truth.

The remainder of our paper proceeds as follows. Section 2 describes our statistical model and estimation procedure. Section 3 describes our results for the equity premium, and extends these results to international data and to characteristic-sorted portfolios. Because we find a larger reduction for small stocks as compared to large stocks, our results suggest that the size premium, as well as the equity premium, may have been a result of an unusual series of shocks. Section 4 describes the intuition for our efficiency results and how these results depend on the parameters of the data generating process and the length of the time series. Section 5 shows the applicability of our procedure under alternative data generating processes, including conditional heteroskedasticity and structural breaks. Section 6 concludes.

2. Statistical model and estimation

This section gives the specifics of our benchmark statistical model (Section 2.1), describes our estimation method (Section 2.2), and our data (Section 2.3).

2.1. Statistical model

Let R_{t+1} denote net returns on an equity index between t and $t+1$, and $R_{f,t+1}$ denote net riskfree returns between t and $t+1$. We let $r_{t+1} = \log(1+R_{t+1}) - \log(1+R_{f,t+1})$. Let x_t denote the log of the dividend-price ratio. We assume

$$r_{t+1} - \mu_r = \beta(x_t - \mu_x) + u_{t+1} \tag{1a}$$

$$x_{t+1} - \mu_x = \theta(x_t - \mu_x) + v_{t+1}, \tag{1b}$$

where, conditional on $(r_1, \ldots, r_t, x_0, \ldots, x_t)$, the vector of shocks $[u_{t+1}, v_{t+1}]^\top$ is normally distributed with zero mean and covariance matrix

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

We assume that the dividend-price ratio follows a stationary process, namely, that $-1 < \theta < 1$; later we discuss the implications of relaxing this assumption. Taking expectations on both sides of (1a) and (1b) implies that μ_r is the unconditional mean of r_t (namely, the equity premium), and μ_x as the unconditional mean of x_t.

The system of equations in (1) is standard in the literature. Indeed, (1a) is equivalent to the ordinary least squares regression that has been a focus of measuring predictability in stock returns for almost 30 years (Keim and Stambaugh, 1986; Fama and French, 1989). We have simply rearranged the parameters so that the mean excess return μ_r appears explicitly. The stationary first-order autoregression for x_t is standard in settings where modeling x_t is necessary, e.g. understanding long-horizon returns or the statistical properties of estimators for β.[2] Indeed, most leading economic models imply that x_t is stationary (e.g. Bansal and Yaron, 2004; Campbell and Cochrane, 1999). A large and sophisticated literature uses this setting to explore the bias and size distortions in estimation of β, treating other parameters, including μ_r, as "nuisance" parameters.[3] Our work differs from this literature in that μ_r is not

[2]See for example Campbell and Viceira (1999), Barberis (2000), Fama and French (2002), Lewellen (2004), Cochrane (2008), van Binsbergen and Koijen (2010).

[3]See for example Bekaert et al. (1997), Campbell and Yogo (2006), Nelson and Kim (1993), and Stambaugh (1999) for discussions on the bias in estimation of β and Cavanagh et al. (1995), Elliott and Stock (1994), Jansson and Moreira (2006), Torous et al. (2004) and Ferson et al. (2003) for discussion of size. Campbell (2006) surveys this literature. There is a connection between estimation of the mean and of the predictive coefficient, in that the bias in β arises from the bias in θ (Stambaugh, 1999), which ultimately arises from the need to estimate μ_x (Andrews, 1993).

a nuisance parameter but rather the focus of our study.

A classic motivation for (1) is the tight theoretical connection between realized returns, expected future returns, and the dividend-price ratio (Campbell and Shiller, 1988). For the purpose of this discussion, let r_t denote the log of the return on the stock market index (rather than the excess return), let p_t denote the log price, and d_t the log dividend. It follows from the definition of a return that

$$r_{t+1} = \log(e^{p_{t+1}-d_{t+1}} + 1) - (p_t - d_t) + d_{t+1} - d_t.$$

Applying a Taylor expansion, as in Campbell (2003), implies

$$r_{t+1} \approx \text{constant } + k(p_{t+1} - d_{t+1}) + d_{t+1} - p_t$$

where $k \in (0, 1)$. Thus, with $x_t = d_t - p_t$, it follows that

$$r_{t+1} - E_t[r_{t+1}] = -k\left(x_{t+1} - E_t[x_t]\right) + d_{t+1} - E_t[d_{t+1}]. \tag{2}$$

Equation 2 establishes that, as a matter of accounting, we would expect that shocks to returns and shocks to the dividend-price ratio to be negatively correlated. That is, $\rho_{uv} < 0$ in the equations above.

By solving these equations forward, Campbell (2003) further derives the present-value identity

$$x_t = \text{constant} + E_t \sum_{j=0}^{\infty} k^j \left(r_{t+1+j} - \Delta d_{t+1+j}\right). \tag{3}$$

Equation 3 provides a second link between the dividend-price ratio and returns, namely, that the dividend-price ratio x_t should pick up variation in future discount rates ($\beta > 0$ in (1a)). Given (3), it follows from (2) that shocks to returns can be expressed as

$$r_{t+1} - E_t r_{t+1} = (E_{t+1} - E_t) \sum_{j=0}^{\infty} k^j \Delta d_{t+1+j} - (E_{t+1} - E_t) \sum_{j=1}^{\infty} k^j r_{t+1+j}. \tag{4}$$

There is a longstanding debate about which term in (4), expected future cash flows or discount rates, is responsible for the volatility of the dividend-price ratio. As we will show, our method is agnostic when it comes to this question. What we will require is the first link described in the paragraph above: persistent variation in the dividend-price ratio (which could be driven either by discount rates or cash flows) that is negatively correlated with realized returns.[4]

2.2. Estimation procedure

We estimate the parameters μ_r, μ_x, β, θ, σ_u^2, σ_v^2 and σ_{uv} by maximum likelihood. The assumption on the shocks implies that, conditional on the first observation x_0, the likelihood function is given by

$$p\left(r_1,\ldots,r_T;x_1,\ldots,x_T|\mu_r,\mu_x,\beta,\theta,\Sigma,x_0\right) = \\ |2\pi\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T}u_t^2 - 2\frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T}u_t v_t + \frac{\sigma_u^2}{|\Sigma|}\sum_{t=1}^{T}v_t^2\right)\right\}. \quad (5)$$

Maximizing this likelihood function is equivalent to running ordinary least squares regression (Davidson and MacKinnon, 1993, Chapter 8). Not surprisingly, maximizing the above requires choosing means and predictive coefficients to minimize the sum of squares of u_t and v_t.

This likelihood function, however, ignores the information contained in the initial draw x_0. For this reason, studies have proposed a likelihood function that incorporates the first observation (Box and Tiao, 1973; Poirier, 1978), assuming that it is a draw from the stationary distribution. In our case, the

[4]These considerations motivate our focus on the dividend-price ratio throughout this manuscript. Moreover, the economic reasons for our effect are easiest seen in a univariate setting. As an empirical matter, adding variables such as the default spread and term spread to (1) has little effect beyond what we find with the dividend-price ratio. See Table D.5 in the Online Appendix.

stationary distribution of x_0 is normal with mean μ_x and variance

$$\sigma_x^2 = \frac{\sigma_v^2}{1-\theta^2},$$

(Hamilton, 1994). The resulting likelihood function is

$$p\left(r_1,\ldots,r_T;x_0,\ldots,x_T|\mu_r,\mu_x,\beta,\theta,\Sigma\right) = \left(2\pi\sigma_x^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{x_0-\mu_x}{\sigma_x}\right)^2\right\}\times$$
$$|2\pi\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T}u_t^2 - 2\frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T}u_t v_t + \frac{\sigma_u^2}{|\Sigma|}\sum_{t=1}^{T}v_t^2\right)\right\}. \quad (6)$$

Likelihood function (5) is often referred to as the conditional likelihood and (6) as the exact likelihood. Papers that makes use of the exact likelihood in the context of return estimation include Stambaugh (1999) and Wachter and Warusawitharana (2009, 2012), who focus on estimation of the predictive coefficient β.[5] In contrast, van Binsbergen and Koijen (2010), who focus on return predictability in a latent-variable context, use the conditional likelihood function (with the assumption of stationarity). Other previous studies have focused on the effect of the exact likelihood on unit root tests (Elliott, 1999; Müller and Elliott, 2003).

We derive the values of μ_r, μ_x, β, θ, σ_u^2, σ_v^2 and σ_{uv} that maximize the likelihood (6) by solving a set of first-order conditions. We give closed-form expressions for each maximum likelihood estimate in Appendix A. Our solution amounts to solving a polynomial for the autoregressive coefficient θ, after which the solution of every other parameter unravels easily. Because our method does not require numerical optimization, it is computationally expedient. We will refer to this procedure as maximum likelihood estimation (MLE) even when we examine cases in which it is mis-specified. We focus

[5] Wachter and Warusawitharana (2009, 2012) use Bayesian methods rather than maximum likelihood.

on a comparison with the most common alternative way of calculating the equity premium, namely the sample average. Note that this sample average would appear as the constant term an OLS regression of returns on a predictor variable that is demeaned using the first $T-1$ observations.

Given that our goal is to estimate μ_r, which is a parameter determining the marginal distribution of returns, why might it be beneficial to jointly estimate a process for returns and for the dividend-price ratio? Here, we give a general answer to this question, and go further into specifics in Section 4. First, a standard result in econometrics says that maximum likelihood, assuming that the specification is correct, provides the most efficient estimates of the parameters, that is, the estimates with the (weakly) smallest asymptotic standard errors (Amemiya, 1985). Furthermore, in large samples, and assuming no mis-specification, introducing more data makes inference more reliable rather than less. Thus the value of μ_r that maximizes the likelihood function (6) should be (asymptotically) more efficient than the sample mean because it is a maximum likelihood estimator and because it incorporates more data than a simpler likelihood function based only on the unconditional distribution of the return r_t.[6]

This reasoning holds asymptotically. Several considerations may work against this reasoning in small samples. First, asymptotic theory says only that maximum likelihood is better (or, technically, at least as good), but the difference may be negligible. Second, even if there is an improvement in asymptotic efficiency for maximum likelihood, it could easily be outweighed in practice by the need to estimate a more complicated system. Finally, estimation of the

[6]The distinction between a multivariate and univariate system calls to mind the distinction between Seemingly Unrelated Regression (SUR) and OLS (Zellner, 1962). As will become clear in what follows, our results do not arise from the use of the multivariate system per se (as Zellner shows, there is no efficiency gain to multivariate estimation when the right-hand-side variables are the same). Rather, the gains arise from the multivariate system in combination with the initial term in the exact likelihood function.

equity premium by the sample mean does not require specification of the predictor process. Mis-specification in the process for dividend-price ratio could outweigh the benefits from maximum likelihood. These questions motivate the analysis that follows.

2.3. Data

In what follows, our market return is defined as the monthly value-weighted return on the NYSE/AMEX/NASDAQ available from CRSP. Using returns with and without dividends, we construct a monthly dividend series. We then follow the standard construction for the dividend-price ratio that eliminates seasonality, namely, we divide a monthly dividend series (constructed by summing over dividend payouts over the current month and previous eleven months) by the price.

We also consider returns on portfolios formed on the basis of size and book-to-market. Again we use value-weighted returns with and without dividends to construct a dividend series for each portfolio. We then construct a dividend-price ratio series for each portfolio in the same manner as for the market portfolio. We also consider dollar returns on international and country-level indices. For each of these, we construct a dividend-price ratio series in the same manner described above. International return data are available from Kenneth French's website. Fama and French (1989) discuss details of the construction of these data.

To form an excess return, we subtract the monthly return on the 30-day Treasury Bill. Given the net return R_t on the equity series and the net Treasury return R_t^f, we take $r_t = \log(1 + R_t) - \log(1 + R_t^f)$.

3. Results

This section describes our main results. Section 3.1 describes the results of maximum likelihood estimation for the aggregate market and compares it with ordinary least squares and sample means. Section 3.2 describes out-of-sample tests of our method. Section 3.3 describes an application to the cross-section of returns and Section 3.4 to international data. Section 3.5 describes results for valuation measures other than the dividend-price ratio.

3.1. Point estimates for the U.S. equity premium

Table 1 reports maximum likelihood estimates of the parameters of our statistical model given in (1) under the heading MLE. We report estimates for the 1927-2011 sample and for the 1953-2011 postwar subsample. In this section we discuss point estimates, and postpone the discussion of standard errors, and the statistical efficiency of the estimates, to Section 3.6. For comparison, we also report the sample average of excess returns and the sample mean of the dividend-price ratio under the heading Sample. For the postwar sample, this sample average is 0.433% in monthly terms, or 5.20% per annum. In contrast, the maximum likelihood estimate of the equity premium is 0.322% monthly, or 3.86% per annum. The annualized difference is 133 basis points. Applying MLE to the 1927–2011 sample yields an estimated mean of 4.69% per annum, 88 basis points lower than the sample average.[7]

Maximum likelihood also implies a different estimate for the mean of the dividend-price ratio than the sample average. The difference is relatively small, however; only 4 basis point in the postwar data, an order of magnitude smaller than the difference in the estimate of the equity premium. Nonetheless, the two results are closely related, as we will discuss in what follows.

[7]When translated to annualized level returns, the per annum estimate falls from 6.37% to 5.06%. See Appendix B.1.

Maximum likelihood gives values for the predictive coefficient β, the autocorrelation θ, and the variance-covariance matrix Σ. We compare these to values of β and θ from traditional OLS forecasting regressions on a constant and on the lagged dividend-price ratio. We report the results for β and θ, as well as the variance-covariance matrix, in Table 1 under the heading OLS. The estimate of the variance-covariance matrix are nearly identical (by definition, the estimates of σ_u and σ_v will be higher under MLE than under OLS; we find no noticeable difference for σ_v and a negligible difference for σ_u). This is not surprising, as volatility is known to be estimated precisely in monthly data. Estimates for the regression coefficient β are noticeably different. In postwar data, maximum likelihood estimates a lower value of β (0.69 vs. 0.83). This lower estimate for β is driven by the (slightly) higher estimate for the autocorrelation coefficient θ (deviations of β and θ from their OLS values go in opposite directions, see Stambaugh (1999)). The result, however, is sample-dependent. In the longer sample, the maximum likelihood estimate for β is higher than the OLS value, and naturally the estimate for θ is lower.

Given the controversy surrounding the parameter β, we next ask how the estimation of predictability affects our results. We repeat maximum likelihood estimation, but restrict β to be zero. That is, we consider

$$r_{t+1} - \mu_r = u_{t+1} \tag{7a}$$

$$x_{t+1} - \mu_x = \theta(x_t - \mu_x) + v_{t+1}, \tag{7b}$$

In what follows, we refer to this as *restricted maximum likelihood*, and use the terminology MLE_0.[8] Table 1 shows, perhaps surprisingly, that the maximum likelihood estimate for the mean return hardly changes. It is in fact slightly lower (0.31% vs. 0.32%) in postwar data, and thus further away from the sample mean. The most notable difference between the two types of estimation is the value for the autocorrelation θ, which is closer to unity under MLE_0.

[8]See Appendix A.2 for more details on our methodology.

Given that the right-hand-side variables of the two equations are no longer the same, it is possible for estimation of the system to yield different results than estimation of each equation separately (Zellner, 1986).[9] Moreover, if the true value of β is equal to zero but the OLS value is positive, realized shocks must be such that the true autocorrelation of x_t is higher than the measured one (Wachter and Warusawitharana, 2015).

The restricted maximum likelihood estimation indicates that return predictability is not driving our results. In fact, it arises because MLE allows us to incorporate information about the stationary distribution of x_t. This information leads us to conclude that shocks to the dividend-price ratio have been negative on average. The negative contemporaneous correlation between shocks to the dividend-price ratio and to returns allows allows this information to be incorporated into the estimation of the return process: shocks to returns have been positive on average and thus some of the measured equity premium is due to good luck. We discuss this intuition in more detail in Section 4.

3.2. Out-of-sample results

While we are using the system (1) to estimate the unconditional mean μ_r, much of the prior literature focuses on estimating the conditional equity premium, namely the forecast for excess stock returns conditional on x_t. Such forecasts have been found to have inferior out-of-sample performance as compared to the sample average (Bossaerts and Hillion, 1999; Welch and Goyal, 2008).[10] This raises the question of whether our unconditional estimates, coming from a conditional model, can outperform the sample average.

[9] The presence of the initial condition in the likelihood function implies that our estimates will not be identical to OLS regardless. However, the effect of the initial condition on estimation of θ is small compared to the effect of restricting β to be zero.

[10] Alternative means of incorporating information can lead some conditional models to outperform, e.g. Campbell and Thompson (2008) and Kelly and Pruitt (2013).

To answer this question, we compute the root-mean-squared-error (RMSE) based on our estimate versus the sample mean. Specifically, for each observation (starting ten years after the start of our sample), we compute both the maximum likelihood estimate and the sample mean using the previous data. We then take the difference between the stock return and this estimate over the following month and square it. Summing these up, dividing by the number of observations, and taking the square root yields the RMSE.

A caveat to this analysis is in order. Given that we are only attempting an unconditional estimate of the mean, the best we could possibly do in terms of RMSE would be the realized unconditional standard deviation of stock returns over the sample. This is what we would find if we could estimate the mean perfectly. That is, the "error" used to compute the RMSE is in fact the variation in stock returns. This variation is quite high, and is likely to be high compared to possible improvements in the unconditional estimate of the mean.

In fact, we find that unlike conditional mean forecasts that incorporate the dividend-price ratio, our unconditional forecasts yield better out-of-sample performance.[11] The difference in the RMSEs between the sample mean and the MLE is 0.011% per month in the postwar period, or 0.132% per annum. We find very similar results for restricted maximum likelihood. Maximum likelihood also outperforms the sample mean over the period beginning in 1927. The difference between the mean-squared errors are significant using the Diebold and Mariano (2002) test.[12] These results suggest that our estimates

[11]The maximum likelihood estimates are also smoother over time. This result is shown in Figure 5 and discussed in Section 4.2.3.

[12]The autocorrelation of the difference in mean-squared errors is very low. Nonetheless, one concern with this test is that our mean estimators use data from the entire previous sample period, and thus exhibit long-run dependence. In theory, one could mitigate this concern by using rolling estimation windows (as recommended by Giacomini and White, 2006). However, to obtain reliable estimates, one would want to have these rolling windows

are not only different from the sample mean, they are also more reliable. We return to this point in Section 3.6, when we evaluate efficiency.

3.3. Characteristic-sorted portfolios

An advantage of our method is its ease and wide applicability: it is not specific to the market portfolio. To illustrate this, we highlight two additional applications, one to characteristic-sorted portfolios (this section) and to international stock returns (the following section).

We first consider portfolios formed by sorting stocks by market equity and then forming portfolios based on quintiles (see Fama and French (1992) for more detail). Panel A of Table 2 shows the resulting sample means (Sample), maximum likelihood estimates (MLE), and restricted maximum likelihood estimates (MLE_0). The Sample row clearly replicates the classic finding of Fama and French (1992): stocks with low market equity of higher average than stocks with high market equity. The difference is an economically significant 0.16% per month.

The next column re-examines this size finding from the perspective of MLE. We repeat our analysis, using the relevant dividend-price ratio series for each quintile (see Section 2.3) for more information. As for the market portfolio, the use of maximum likelihood significantly reduces the estimated mean on each portfolio. Again, replicating our results for the market portfolio, MLE and MLE_0 consistently lead to lower RMSE in out-of-sample tests across the quintiles.

While the change to the quintiles is all in the same direction (namely, down), the magnitude of the effect differs substantially between the quintiles. The lowest quintile (with the smallest stocks) exhibits the greatest reduction: around 23 basis points. The largest stocks exhibit a reduction of less than one

be large, and so, practically speaking, the long-run dependence problem would still be there.

basis point. As the last column shows, the resulting size premium therefore all but disappears (it is a mere 3 basis points) when MLE is used. Running restricted MLE leads to a similar, and in fact slightly larger, reduction.

Panel B of Table 2 shows analogous results for portfolios formed on the ratio of book equity to market equity. Again, the first row shows sample means, and replicates the result of Fama and French (1992) that stocks with a low ratio of book equity to market equity (growth firms) have substantially lower returns than stocks with a high ratio of book equity to market equity (value firms). The difference is 0.32% per month. Repeating MLE and MLE_0 (again, we construct a dividend-price ratio series for each quintile), we find a reduction in the mean estimate for all portfolios and an improved RMSE. However, unlike for size, there appears to be no relation between the book-to-market ratio and the magnitude of the reduction, leading the value premium, as estimated over this sample, to be largely unchanged.

One implication of our findings for size portfolios is the estimation of the price of risk corresponding to the size factor. As interpreted by Fama and French (1993), the return differential on the small-minus-big portfolio represents not so much an anomaly but a return for bearing risk. Our results suggest that this risk premium is smaller than previously believed. More broadly, one could apply our results to the estimation of risk premiums due to cross-sectional factors. If one is deriving these risk premiums from a two-pass regression approach (the first pass would be to estimate expected returns and betas, the second to estimate the risk premiums due to the factors), our method could be used to estimate the risk premiums in the first pass. This would imply greater precision in the estimation of the risk premiums in the second stage.

3.4. International stock returns

A natural question is whether the reduction in the equity premium is U.S. specific, or a feature of global financial markets. We first consider return data on regional indices (which begin in 1976), and report the results in Table 3. We find that a value-weighted index meant to proxy for the world portfolio falls by nearly half, from 0.36% per month to 0.19% per month. The Asia index falls by even more: 0.26% per month to 0.12%. However, the EU index (with the UK included) is affected by comparatively little: the premium falls from 0.42% to 0.33%. Clearly our results are not specific to the U.S. market. These dramatic results reflect the power of our approach in samples that are relatively short.

When we apply our estimation to country-level stock-market indices, some interesting differences emerge. Results are reported in Table 4. For some countries, nearly all the expected return appears to be due to luck (for example, Japan and Italy). We also find that our measure concludes that "bad luck" has caused some returns to be understated, for example, Denmark and Spain. The findings for both regional and country-level data are consistent across MLE and MLE_0 methods, indicating that these findings are not driven by return predictability.[13]

3.5. Alternative valuation measures

The discussion in Section 2.1 indicates that the dividend-price ratio has a special role in our maximum-likelihood analysis. Because of the present-value identity, we would expect a high correlation between shocks to this series and shocks to realized returns. However, the determinants of firm dividend policy

[13]Given the short data sample available, RMSEs are particularly noisy. However, we find that, on average, MLE has a lower RMSE than the sample mean, both for the regional indices and country-level data.

have been subject to long debate. Moreover, Fama and French (2001) show that the tendency to pay dividends may not be stable over time.

The arbitrariness of dividend payments, and their apparent instability, need not affect our results. The return measure r_t and the dividend payout correspond to what would actually be received by an investor who holds the CRSP value-weighted portfolio, and thus the present-value identity of Section 2.1 is valid. The reasons considered by Fama and French (2001) for reductions in dividend payments are consistent with a stationary dividend-price ratio combined with long-term fluctuations in expected dividend growth; note that our method allows fluctuations in the dividend-price ratio to be due to changes in growth expectations as well as discount rates. Alternatively, perhaps the dividend-price ratio is subject to a structural break. We confront this possibility explicitly in Section 5.3.

We can confirm that our results are not due to unusual characteristics of the dividend series by considering other valuation ratios such as book-to-market or earnings-to-price. Data on total book value and total market value for the S&P 500 is available from Global Financial Database. We take the log of the book-to-market ratio constructed from these data, and apply our maximum likelihood estimator, first using the value-weighted CRSP return, and then the changes in price on the S&P 500. These data are monthly and begin in 1977. We show results in Table 5. Maximum likelihood implies an equity premium of 0.30% per month on the CRSP portfolio, as compared with a sample mean of 0.43% over the same period. These are stronger results than we find for our benchmark estimation (the results for restricted maximum likelihood and for S&P 500 returns are stronger still). For earnings-to-price, we use the CAPE measure proposed by Shiller (2000) that eliminates short-term fluctuations in earnings (data are available from Robert Shiller's website). Because this series has a much lower correlation with CRSP returns, the improvements from maximum likelihood are likely to be smaller (see Section 4). However, even the

earnings-to-price ratio implies a reduction from 0.43 to 0.38 per month. Thus the book-to-market and earnings-to-price series confirm our findings from the dividend-price ratio.[14]

3.6. Efficiency

So far we have shown that MLE gives different estimates for the equity premium than the sample average and that they are more reliable as measured by the root-mean-squared error. We now conduct a formal statistical comparison of the methods. Namely, we ask whether our procedure reduces estimation noise in finite samples.

We simulate 10,000 samples of length equal to the postwar data. We draw returns and predictor variable observations from (1), setting parameter values equal to their maximum likelihood estimates (in what follows we refer to this as the benchmark simulation). For each sample, we initialize x with a draw from the stationary distribution. We then calculate sample averages, OLS estimates and maximum likelihood estimates for each sample path, generating a distribution of these estimates over the 10,000 paths.[15]

[14]We find similarly large reductions in the mean return when we adjust the dividend series for repurchases as in Boudoukh et al. (2007), as reported in Table D.6 of the Online Appendix (annual data are available from the website of Michael Roberts until 2003). As these authors note, however, to be consistent one should adjust for stock market issuance as well as repurchases (and ideally for repayment and issuance of debt as well; see Larrain and Yogo, 2008). Stock-market issuance relative to market value has essentially zero correlation with realized returns. Shocks to the issuance-adjusted series convey relatively little information about shocks to returns, at least as measured with our current methods. This information transmission is crucial for our method to deliver large efficiency gains relative to the sample mean.

[15]In every sample, both actual and artificial, we have been able to find a unique solution to the first order conditions such that θ is real and between -1 and 1. Given this value for θ, there is a unique solution for the other parameters. The distribution for these values is shown in Figure D.1. See Appendix A for further discussion of the polynomial for θ.

Table 6 reports means, standard deviations, and the 5th, 50th, and 95th percentile values. Note that these statistics refer to the sampling distribution. Therefore, the standard deviations should be interpreted as standard errors for the corresponding estimates in Table 1. Table 6 shows that, while the sample average of the excess return has a standard error of 0.089 (in monthly percentage terms), the maximum likelihood estimate has a standard error of only 0.050.

Besides lower standard errors, the maximum likelihood estimates also have a tighter distribution. For example, the 95th percentile value for the sample mean of returns is 0.47, while the 95th percentile value for the maximum likelihood estimate is 0.40 (in monthly terms, the value of the maximum likelihood estimate is 0.32). The 5th percentile is 0.18 for the sample average but 0.24 for the maximum likelihood estimate. This tighter distribution can clearly be seen in Figure 1, which shows the distribution of the maximum likelihood estimates is visibly more concentrated around the true value of the equity premium, and that the tails of this distribution fall well under the tails of the distribution of sample means.

Table 6 also shows that the maximum likelihood estimate of the mean of the predictor has a lower standard error and tighter confidence intervals than the sample average, though the difference is much less pronounced. Similarly, the maximum likelihood estimate of the regression coefficient β also has a smaller standard error and tighter confidence bands than the OLS estimate, though again, the differences for these parameters between MLE and OLS are not large. The results in this table show that no other parameters are subject to the same dramatic improvement in efficiency as the mean return. This is in part due to the fact that estimation of first moments in a time series is more difficult than that of second moments (Merton, 1980). It is also due to the relatively high volatility of shocks to returns, as we discuss in Section 4.

The results in Table 6 also illustrate the bias in the OLS estimate of the

predictive coefficient β (Stambaugh, 1999). While the data generating process assumes a β value of 0.69, the mean OLS value from the simulated samples is 1.28. That is, OLS estimates the predictive coefficient to be much higher than the true value, and thus the predictive relation to be stronger. The bias in the predictive coefficient is associated with bias in the autoregressive coefficient on the dividend-price ratio. The true value of θ in the simulated data is 0.993, but the mean OLS value is 0.987. Maximum likelihood reduces the bias somewhat: the mean maximum likelihood estimate of β is 1.24 as opposed to 1.28, but it does not eliminate it. Note that the estimates of the equity premium are not biased; the mean for both maximum likelihood and the sample average is close to the population value.

These results suggest that 0.69 is probably not a good estimate of β, and likewise, 0.993 is likely not to be a good estimate of θ. Does the superior performance of maximum likelihood continue to hold if these estimates are corrected for bias? We turn to this question next. We repeat the exercise described above, but instead of using the maximum likelihood estimates, we adjust the values of β and θ so that the mean computed across the simulated samples matches the observed value in the data. The results are given in Panel B. This adjustment lowers β and increases θ, but does not change the maximum likelihood estimate of the equity premium. If anything, adjusting for biases shows that we are being conservative in how much more efficient our method of estimating the equity premium is in comparison to using the sample average. The sample average has a standard deviation of 0.138, while the standard deviation of the maximum likelihood estimate is 0.072. After accounting for biases, maximum likelihood gives an equity premium estimate with standard deviation that is about half of the standard deviation of the sample mean excess return. We will refer to this as our benchmark case with bias-correction.

These results show that our efficiency gains continue to hold after correcting

for bias in the predictive coefficient. In fact, the result that maximum likelihood is more efficient is quite robust. As we show in the Online Appendix, it holds when we assume fat-tailed shocks (Table D.1), and when we use the OLS estimates rather than the maximum likelihood estimates (Table D.2). While longer data series naturally produce a smaller improvement, we still see an economically significant reduction in standard errors in Monte Carlo experiments designed to match the sample beginning in 1927 (Table D.3). Restricted maximum likelihood is also more efficient than the sample mean (Table D.4). In Section 5 we consider substantial departures from our data generating process, as well the potential for structural breaks, and continue to find large efficiency gains. Finally, we can also see the efficiency gains in asymptotic standard errors. Because the likelihood function is available in closed form, we can calculate well-behaved asymptotic standard errors as explained in Appendix A.4. We report these in Table 7.[16] The asymptotic standard error for maximum likelihood is 0.054, almost identical to its finite-sample counterpart, 0.50. The standard error on the sample mean is larger than its finite-sample counterpart: 0.114 as compared to 0.089. This implies even greater efficiency gains from maximum likelihood when evaluated using traditional asymptotics. In the sections that follow, we explain the source of this efficiency gain, and why it is so robust to variations in our assumptions.

4. Discussion

The previous section showed that maximum likelihood is a more efficient estimator than the sample mean. This efficiency is of economic consequence: not

[16]We calculate standard errors for the sample averages taking into account the autocorrelation structure in the data. Given (1), the variance of the sample mean of returns and of x_t are available in closed form (see Appendix C.2). We substitute in the series of shocks u_t and v_t from maximum likelihood estimation, and bias-corrected values for β and θ.

only does maximum likelihood give a substantially different estimate of the equity premium, it gives a more reliable one. In Section 4.1 we discuss the reason why maximum likelihood is more efficient. In Section 4.2 we discuss the properties of the time series that determine these efficiency gains. This is useful for researchers in determining when our method is likely to be of the greatest value.

4.1. Source of the gain in efficiency

What determines the difference between the maximum likelihood estimate of the equity premium and the sample average of excess returns? Let $\hat{\mu}_r$ denote the maximum likelihood estimate of the equity premium and $\hat{\mu}_x$ the maximum likelihood estimate of the mean of the dividend-price ratio. Given these estimates, we can define a time series of shocks $\hat{u}_t$ and $\hat{v}_t$ as follows:

$$\hat{u}_t = r_t - \hat{\mu}_r - \hat{\beta}(x_{t-1} - \hat{\mu}_x) \tag{8a}$$

$$\hat{v}_t = x_t - \hat{\mu}_x - \hat{\theta}(x_{t-1} - \hat{\mu}_x). \tag{8b}$$

By definition, then,

$$\hat{\mu}_r = \frac{1}{T}\sum_{t=1}^{T} r_t - \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t - \hat{\beta}\frac{1}{T}\sum_{t=1}^{T}(x_{t-1} - \hat{\mu}_x). \tag{9}$$

As (9) shows, there are two reasons why the maximum likelihood estimate of the mean, $\hat{\mu}_r$, might differ from the sample mean $\frac{1}{T}\sum_{t=1}^{T} r_t$. The first is that the shocks $\hat{u}_t$ may not average to zero over the sample. The second, which depends on return predictability, is that the average value of x_t might differ from $\hat{\mu}_x$.

It turns out that only the first of these effects is quantitatively important for our sample. For the period January 1953 to December 2001, the sample average $\frac{1}{T}\sum_{t=1}^{T} \hat{u}_t$ is equal to 0.1382% per month, while $\hat{\beta}\frac{1}{T}\sum_{t=1}^{T}(x_{t-1} - \hat{\mu}_x)$ is -0.0278% per month. The difference in the maximum likelihood estimate

and the sample mean thus ultimately comes down to the interpretation of the shocks $\hat{u}_t$. To understand the behavior of these shocks, we will argue it is necessary to understand the behavior of the shocks $\hat{v}_t$. And, to understand $\hat{v}_t$, it is necessary to understand why the maximum likelihood estimate of the mean of x_t differs from the sample mean.

4.1.1. Estimation of the mean of the predictor variable

To build intuition, we consider a simpler problem in which the true value of the autocorrelation coefficient θ is known. We show in Appendix A that the first-order condition in the exact likelihood function with respect to μ_x implies

$$\hat{\mu}_x = \frac{(1+\theta)}{1+\theta+(1-\theta)T} x_0 + \frac{1}{(1+\theta)+(1-\theta)T} \sum_{t=1}^{T} (x_t - \theta x_{t-1}). \tag{10}$$

We can rearrange (1b) as follows:

$$x_{t+1} - \theta x_t = (1-\theta)\mu_x + v_{t+1}.$$

Summing over t and solving for μ_x implies that

$$\mu_x = \frac{1}{1-\theta} \frac{1}{T} \sum_{t=1}^{T} (x_t - \theta x_{t-1}) - \frac{1}{T(1-\theta)} \sum_{t=1}^{T} v_t, \tag{11}$$

where the shocks v_t are defined using the mean μ_x and the autocorrelation θ.

Consider the conditional maximum likelihood estimate of μ_x, the estimate that arises from maximizing the conditional likelihood (5). We will call this $\hat{\mu}_x^c$. Note that this is also equal to the OLS estimate of μ_x, which arises from estimating the intercept $(1-\theta)\mu_x$ in the regression equation

$$x_{t+1} = (1-\theta)\mu_x + \theta x_t + v_{t+1}$$

and dividing by $1-\theta$. The conditional maximum likelihood estimate of μ_x is determined by the requirement that the shocks v_t average to zero. Therefore,

it follows from (11) that

$$\hat{\mu}_x^c = \frac{1}{1-\theta}\frac{1}{T}\sum_{t=1}^{T}(x_t - \theta x_{t-1}).$$

Substituting back into (10) implies

$$\hat{\mu}_x = \frac{(1+\theta)}{1+\theta+(1-\theta)T}x_0 + \frac{(1-\theta)T}{(1+\theta)+(1-\theta)T}\hat{\mu}_x^c.$$

Multiplying and dividing by $1-\theta$ implies a more intuitive formula:

$$\hat{\mu}_x = \frac{1-\theta^2}{1-\theta^2+(1-\theta)^2T}x_0 + \frac{(1-\theta)^2T}{1-\theta^2+(1-\theta)^2T}\hat{\mu}_x^c. \tag{12}$$

Equation 12 shows that the exact maximum likelihood estimate is a weighted average of the first observation and the conditional maximum likelihood estimate. The weights are determined by the precision of each estimate. Recall that

$$x_0 \sim \mathcal{N}\left(0, \frac{\sigma_v^2}{1-\theta^2}\right).$$

Also, because the shocks v_t are independent, we have that

$$\frac{1}{T(1-\theta)}\sum_{t=1}^{T} v_t \sim \mathcal{N}\left(0, \frac{\sigma_v^2}{T(1-\theta)^2}\right).$$

Therefore $T(1-\theta)^2$ can be viewed as proportional to the precision of the conditional maximum likelihood estimate, just as $1-\theta^2$ can be viewed as proportional to the precision of x_0. Note that when $\theta = 0$, there is no persistence and the weight on x_0 is $1/(T+1)$, its appropriate weight if all the observations were independent. At the other extreme, as θ approaches 1, less and less information is conveyed by the shocks v_t and the "estimate" of $\hat{\mu}_x$ approaches x_0.[17]

[17] We cannot use (12) to obtain our maximum likelihood estimate because θ is not known (more precisely, the conditional and exact maximum likelihood estimates of θ will differ). Because of the need to estimate θ, the conditional likelihood estimator for μ_x is much less efficient than the exact likelihood estimator; a fact that is not apparent from these equations.

While (12) rests on the assumption that θ is known, we can nevertheless use it to qualitatively understand the effect of including the first observation. Because of the information contained in x_0, we can conclude that the last T observations of the predictor variable are not representative of values of the predictor variable in population. These values are lower, on average, than they would be in a representative sample. It follows that the predictor variable must have declined over the sample period. Thus the shocks v_t do not average to zero, as OLS (conditional maximum likelihood) would imply, but rather, they average to a negative value.

Figure 2 shows the historical time series of the dividend-price ratio, with the starting value in bold, and a horizontal line representing the mean. Given the appearance of this figure, the conclusion that the dividend-price ratio has been subject to shocks that are negative on average does not seem surprising.

4.1.2. Estimation of the equity premium

We now return to the problem of estimating the equity premium. Equation 9 shows that the average shock $\frac{1}{T}\sum_{t=1}^{T}\hat{u}_t$ plays an important role in explaining the difference between the maximum likelihood estimate of the equity premium and the sample mean return. In traditional OLS estimation, these shocks must, by definition, average to zero. When the shocks are computed using the (exact) maximum likelihood estimate, however, they may not.

To understand the properties of the average shocks to returns, we note that the first-order condition for estimation of $\hat{\mu}_r$ implies

$$\frac{1}{T}\sum_{t=1}^{T}\hat{u}_t = \frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2}\frac{1}{T}\sum_{t=1}^{T}\hat{v}_t. \tag{13}$$

This is analogous to a result of Stambaugh (1999), in which the averages of the error terms are replaced by the deviation of β and of θ from the true means. Equation 13 implies a connection between the average value of the shocks to

the predictor variable and the average value of the shocks to returns. As the previous section shows, MLE implies that the average shock to the predictor variable is negative in our sample. Because shocks to returns are negatively correlated with shocks to the predictor variable, the average shock to returns is positive.[18] Note that this result operates purely through the correlation of the shocks, and is not related to predictability.

Based on this intuition, we can label the terms in (9) as follows:

$$\hat{\mu}_r = \frac{1}{T}\sum_{t=1}^{T} r_t \quad - \quad \underbrace{\frac{1}{T}\sum_{t=1}^{T} \hat{u}_t}_{\text{Correlated shock term}} \quad - \quad \underbrace{\hat{\beta}\frac{1}{T}\sum_{t=1}^{T}(x_{t-1} - \hat{\mu}_x)}_{\text{Predictability term}}. \tag{14}$$

As discussed above, the correlated shock term accounts for more than 100% of the difference between the sample mean and the maximum likelihood estimate of the equity premium, and is an order of magnitude larger than the predictability term. Our argument above can be extended to show why these terms tend to have opposite signs. When the correlated shock term is positive (as is the case in our data), shocks to the dividend-price ratio must be negative over the sample. The estimated mean of the predictor variable will therefore be above the sample mean, and the predictability term will be negative. Figure D.2 shows that indeed these terms tend to have opposite signs in the simulated data.[19]

[18] This point is related to the result that longer time series can help estimate parameters determined by shorter time series, as long as the shocks are correlated (Stambaugh, 1997; Singleton, 2006; Lynch and Wachter, 2013). Here, the time series for the predictor is slightly longer than the time series of the return. Despite the small difference in the lengths of the data, the structure of the problem implies that the effect of including the full predictor variable series is very strong.

[19] There is a small opposing effect on the sign of the predictability term. Note that the sample mean in this term only sums over the first $T-1$ observations. If the predictor has been falling over the sample, this partial sum will lie above the sample mean, though probably below the maximum likelihood estimate of the mean.

This section has explained the difference between the sample mean and the maximum likelihood estimate of the equity premium by appealing to the difference between the sample mean and the maximum likelihood estimate of the mean of the predictor variable. However, Table 1 shows that the difference between the sample mean of excess returns and the maximum likelihood estimate of the equity premium is many times that of the difference between the two estimates of the mean of the predictor variable. Moreover, Table 6 shows that the difference in efficiency for returns is also much greater than the difference in efficiency for the predictor variable. How is it then that the difference in the estimates for the mean of the predictor variable could be driving the results? Equation 13 offers an explanation. Shocks to returns are far more volatile than shocks to the predictor variable. The term $\hat{\sigma}_{uv}/\hat{\sigma}_v^2$ is about -100 in the data. What seems like only a small increase in information concerning the shocks to the predictor variable translates to quite a lot of information concerning returns.

4.1.3. Conditional maximum likelihood

In the previous sections, we compare the results from maximizing the exact likelihood function (6) with sample means. We can also compare our results to conditional maximum likelihood estimates, namely the parameter values that maximize the conditional likelihood function (5). Conditional maximum likelihood gives identical results to OLS for the regression parameters β, θ, and the variance-covariance matrix Σ. Based on this result, one might think that the conditional MLEs of μ_r and μ_x would equal the sample means of r_t and x_t. However, they do not.

Consider first the estimation of μ_x. The conditional maximum likelihood estimate for the mean of the log dividend-price ratio is -3.67. This is below the sample mean of -3.55. In contrast, the exact maximum likelihood estimate is -3.50. This wedge between the conditional maximum likelihood estimate and

the sample mean creates a wedge between the conditional maximum likelihood estimate of μ_r and the corresponding sample mean, but in a very different way than for exact maximum likelihood estimation.

To see how the estimation of μ_x affects μ_r in the conditional case, consider (14), which must hold for any estimator of μ_r because it relies only on (1a). A condition of conditional maximum likelihood is that the shocks are on average equal to zero (recall the equivalence with OLS); thus the correlated shock term in (14) disappears. The entire difference between the conditional MLE of μ_r (we will call this $\hat{\mu}_r^c$) and the sample mean of returns is therefore due to return predictability. Because the conditional MLE $\hat{\mu}_x^c$ is far below the sample mean, the predictability term in (14) is positive and large. It follows that, like its exact counterpart, $\hat{\mu}_r^c$ is below the sample mean (it is equal to 0.31 in postwar data). Intuitively, if the dividend-price ratio has been abnormally high in the sample, and if returns have a component that is based on this value, then returns, too, will have been abnormally high.

Thus conditional and exact maximum likelihood estimation are very different. For conditional maximum likelihood, the finding of the lower equity premium depends entirely on stock return predictability; bias-correcting β substantially reduces this result and restricting β to equal zero eliminates it (in this case the equity premium simply equals the sample average excess return). In contrast, for exact MLE, the effect of predictability is small and in the opposite direction. The difference in the estimators for μ_r stems from differences in the estimators for μ_x. Exact maximum likelihood uses information from the level of the series. Conditional maximum likelihood, however, solves

$$\hat{\mu}_x^c = \frac{1}{1-\hat{\theta}^c} \sum_{t=1}^{T} (x_t - \hat{\theta}^c x_{t-1}),$$

Conditional maximum likelihood thus attempts to identify the mean of x_t from its drift over the course of the sample. It divides these tiny increments by another tiny value: $1-\hat{\theta}^c$. The resulting estimates of μ_x are highly unstable.

In simulated data, $\hat{\theta}^c$ falls outside the unit circle in a non-trivial number of sample paths; for these paths the estimate of μ_x, and hence μ_r, fails to exist. In contrast, exact maximum likelihood always returns an estimate of θ that is within the unit circle. That is, the performance of conditional likelihood in estimating the sample mean is sufficiently poor that we cannot evaluate it within our Monte Carlo framework.[20]

4.2. Properties of the maximum likelihood estimator

In this section we investigate how the improvement in precision from maximum likelihood depends on the persistence of x_t, the degree to which stock returns are predictable, the correlation between the shocks to x_t and the shocks to returns, and the length of the time series. Besides giving insight into the mechanism behind the improvement, this section illuminates the practical situations where our method is most useful.

As in Section 3.6, our main tool is Monte Carlo simulations. We calculate the standard deviation of our estimators across simulated sample paths. These standard deviations correspond to finite-sample standard errors. An exception is when we consider the length of the time series; in this case we also show how the estimates of the equity premium change over time in the historical data.

4.2.1. Variance of the estimator as a function of the persistence

The theoretical discussion in the previous section suggests that the persistence θ is an important determinant of the increase in efficiency from maximum

[20] One way around the stationarity problem is to force θ to be less than 1. This is most easily accomplished in a Bayesian setting with a prior on θ (for maximum likelihood, one could define a boundary, but such a boundary would have to be a finite distance from one and would therefore be arbitrary). Wachter and Warusawitharana (2015) demonstrate the instability of conditional estimates of μ_x and μ_r in a Bayesian setting.

likelihood. Figure 3 shows the standard deviation of estimators of the mean of the predictor variable (μ_x) in Panel A and of estimators of the equity premium (μ_r) in Panel B as functions of θ. Other parameters are set equal to their benchmark values, adjusted for bias in the case of β. For each value of θ, we simulate 10,000 samples.

Panel A shows that the standard deviation of both the sample mean and MLE of μ_x are increasing in θ. This is not surprising; holding all else equal, an increase in the persistence of θ makes the observations on the predictor variable more alike, thus decreasing their information content. The standard deviation of the sample mean is larger than the standard deviation of the maximum likelihood estimate, indicating that our results above do not depend on a specific value of θ. Moreover, the improvement in efficiency increases as θ grows larger. Consistent with the results in Table 6, the size of the improvement is small.

Panel B shows the standard deviation of estimators of μ_r. In contrast to the case of μ_x, the relation between the standard deviation and θ is non-monotonic for both the sample mean of excess returns and the maximum likelihood estimate of the equity premium. For values of θ below about 0.998, the standard deviations of the estimates are decreasing in θ, while for values of θ above this number they are increasing. This result is surprising given the result in Panel A. As θ increases, any given sample contains less information about the predictor variable, and thus about returns. One might expect that the standard deviation of estimators of the mean return would follow the same pattern as in Panel A. Indeed, this is the case for part of the parameter space, namely when the persistence of the predictor variable is very close to one.

However, an increase in θ has two opposing effects on the variance of the estimators of the equity premium. On the one hand, an increase in θ decreases the information content of the predictor variable series, and thus of the return series, as described above. On the other hand, for a given β, an

increase in θ raises the R^2 in the return regression. Because innovations to the predictable part of returns are negatively correlated with innovations to the unpredictable part of returns, an increase in θ increases mean reversion (this can be seen directly from the expressions for the autocovariance of returns in Appendix C.1).

This increase in mean reversion has consequences for estimation of the equity premium. Intuitively, if in a given sample there is a sequence of unusually high returns, this will tend to be followed by unusually low returns. Thus a sequence of unusually high observations or unusually low observations are less likely to dominate in any given sample, and so the sample average will be more stable than it would be if returns were iid (see Appendix C.2). Because the sample mean is simply the scaled long-horizon return, our result is related to the fact that mean reversion reduces the variability of long-horizon returns relative to short-horizon returns. For θ sufficiently large, the reduction in information from the greater autocorrelation does dominate the effect of mean-reversion, and the variance of both the sample mean and the maximum likelihood estimate increase. In the limit as θ approaches one, returns become non-stationary and the sample mean has infinite variance.

Panel B of Figure 3 also shows that MLE is more efficient than the sample mean for any value of θ. The benefit of using maximum likelihood increases with θ. Indeed, while the standard deviation of the sample mean falls from 0.14 to 0.12 as θ goes from 0.980 to 0.995, the maximum likelihood estimate falls further, from 0.14 to 0.06. It appears that the benefits from mean reversion and from maximum likelihood reinforce each other.

4.2.2. The role of predictability and of correlated shocks

The previous section established the importance of the persistence of the dividend-price ratio in the precision gains from maximum likelihood. In this section we focus on the two aspects of joint return and dividend-price ratio

process that affect how information about the distribution of the dividend-price ratio affects inference concerning returns: the predictive coefficient β and the correlation of the shocks ρ_{uv}.

We first consider the role of predictability. As (9) shows, the difference between the maximum likelihood estimator can be decomposed into a term originating from non-zero shocks, and a term originating from predictability. More than 100% of our result comes from the correlated shock term; in other words the predictability term works against us. Without the predictability term, our equity premium would be 0.29% per month rather than 0.32%.

This result is not surprising given that the intuition in Section 4.1 points to negative ρ_{uv} rather than positive β as the source of our gains. If this is correct, we should be able to document efficiency gains in simulations where the predictive coefficient is reduced or eliminated entirely. Indeed, Table 6 shows that if we bias-correct β and θ, the efficiency gains are even larger than when parameters are set to the maximum likelihood estimates. In this section, we take this analysis a step further, and set β exactly to zero. We repeat the exercise from Section 4.2.1, calculating the standard deviation of the estimates across different values of θ. When we repeat the estimation, we do not impose $\beta = 0$, which will work against us in finding efficiency gains.

Panel C of Figure 3 shows the results. First, because returns are iid, the standard deviation of the sample mean is independent of θ and is a horizontal line on the graph. The standard deviation of the maximum likelihood estimate is, however, decreasing in θ. As θ increases, the information contained in the first data point carries more weight. Thus the estimator is better able to identify the average sign of the shocks to the dividend-price ratio and thus to expected returns. Consider, for example, an autocorrelation of 0.998 (the bias-corrected value in Panel B of Table 6). As Panel C shows, the standard deviation of the MLE estimator is 0.12 while the standard deviation of the sample mean is 0.17, or nearly 50% greater. Thus neither the reduction in the

equity premium that we observe in the historical sample, nor the efficiency of the maximum likelihood estimator depend on the predictability of returns.

So far we have shown how changes in the persistence, and changes in the predictability of returns impact the efficiency of our estimates. In particular, the efficiency of our estimates does not depend on return predictability. On what, then, does it depend? The above discussion suggests that it depends, critically, on the correlation between shocks to the dividend-price ratio and to returns, because this is how the information from the dividend-price ratio regression finds its way into the return regression. We look at this issue specifically in Panel D of Figure 3, where we set the correlation between the shocks to equal zero. In this figure, returns are no longer iid, which explains why the standard deviation of the sample mean estimate rises as θ increases. On other hand, though there is return predictability, the lack of correlation implies that there is no mean reversion in returns, so the increase is monotonic, as opposed to what we saw in Panel B.[21] Most importantly, this figure shows zero, or negligible, efficiency improvements from MLE. In fact, for all but extremely high values of θ, MLE performs very slightly worse than the sample mean, perhaps because it relies on biased estimates of predictability.[22] This exercise has little empirical relevance as the correlation between returns and the dividend-price ratio is reliably estimated to be strongly negative. Nonetheless, it is a stark illustration of the conditions under which our efficiency gains break down.

[21]However, if the equity premium were indeed varying over time, one would expect return innovations to be negatively correlated with realized returns (Pastor and Stambaugh, 2009).

[22]Though the data generating process assumes bias-corrected estimates, MLE will still find values of β that are high relative to the values specified in the simulation. This will hurt its finite-sample performance.

4.2.3. Sample length and the difference between the estimators

Because both the maximum likelihood and the sample mean are consistent estimators for the equity premium, they should converge to the true equity premium as the sample size goes to infinity. The central limit theorem states that the standard deviation of both sampling distributions should fall approximately at the rate $\sqrt{T}$, where T is the sample size. However, as with all asymptotic arguments, there may be practical considerations (such as the difficulty of maximizing a nonlinear function) why this might not hold.

We can evaluate the convergence using Monte Carlo simulations as in the previous section, except that here we vary the sample size rather than the parameters of the data generating process. Figure 4 shows the standard deviation of the maximum likelihood estimates, the restricted maximum likelihood estimates, and the sample mean, shown on a log scale (because of the $\sqrt{T}$-convergence). Standard deviations the three estimators decline approximately linearly up to sample sizes of about two hundred, with the intercepts on the two maximum likelihood estimators lying far below that of the sample mean. This figure implies that one would expect to see the greatest (absolute) difference in standard errors in small samples. The ratios of the standard errors should remain roughly constant, however, even for sample sizes that are quite large. Beyond sample sizes of two hundred, the standard deviation for the sample mean slopes downward at a higher rate. Nonetheless, maximum likelihood is still clearly more efficient, even for samples as large as 1000.

We can also see the convergence by estimating the equity premium in our historical data at each point in time. Every month, we compute the sample mean, the maximum likelihood estimate, and the restricted maximum likelihood estimate using the data beginning in January 1953 and continuing up until that month. The results are shown in Panel A Figure 5. The estimates are quite noisy at the beginning of the sample when only a few years of data

are used, but they quickly become smoother. This is especially the case for the maximum likelihood estimates (maximum likelihood and restricted maximum likelihood, are nearly identical through the entire sample) than for the sample mean. In fact, this figure shows that maximum likelihood is far more stable than the sample mean throughout the period, and that the sample mean appears to converge (slowly) to the maximum likelihood value.[23] While the improvement offered by maximum likelihood is significant given the full sample of data, it is even more substantial when only a subset of the data are used.

While Panel A Figure 5 illustrates the slow convergence of the sample mean to the maximum likelihood estimate, the figure also shows substantial short-term variation. For example, the estimators give very similar values in the late 1970s and early 1980s, but the values diverge again in the late 1990s. What drives these differences? Panel B subtracts the sample mean from the maximum likelihood estimate, and multiplies the value of $\sqrt{T}$. The similarity between this figure and the time series of the log dividend-price ratio shown in Figure 2 is clear. That is, it is the behavior of the dividend-price ratio, and more precisely, its difference at the end of the sample from its initial value, that largely determines whether the difference between the sample mean and maximum likelihood is large or small. This result is not surprising given the discussion in Section 4.1, which traces the difference in the estimators to the sign of the average shock to the dividend-price ratio over the sample period. For example, in the late 1970s and early 1980s, the dividend-price ratio was close to its value in 1953, and so the maximum likelihood estimate and the sample mean were quite close to each other. In the 1990s, the dividend-price ratio had diverged far from its value and the maximum likelihood estimate

[23]This reduction in noise also occurs for the estimates of β and θ as well. Namely, the similarity reported for β and θ hold for the full sample, they are noticeably different when, say, only 20 or 30 years of data are available.

and the sample mean again were far apart. During this period,, the sample mean increased substantially relative to its value in the 1970s and 1980s. The maximum likelihood estimate did not, interpreting (in retrospect correctly), the higher return observations during the 1990s as an unusual series of shocks.

To summarize the results of this section: we have shown that the efficiency gains of maximum likelihood are greatest when the variable x_t is persistent, and when its shocks are correlated with the shocks to returns. Predictability plays only a minor role in that it reinforcing the benefits of persistence. The method also delivers its greatest improvement when the sample size is relatively short, though there remains significant improvement for sample lengths many times that usually available for financial time series.

5. Estimation under alternative data generating processes

This section shows the applicability of our procedure under alternative data generating processes. Section 5.1 shows how to adapt our procedure to capture conditional heteroskedasticity in returns and in the predictor variable. Section 5.1 and Section 5.2 consider the performance of our benchmark procedure when confronted with data generating processes that depart from the stationary homoskedastic case in important ways. Our aim is to map out cases where mis-specification overwhelms the gains from introducing data on the dividend-price ratio, and when it does not. Finally, Section 5.3 analysis the consequences of structural breaks for our results.

5.1. Conditional heteroskedasticity

It is well known that stock returns exhibit time-varying volatility (French et al., 1987; Schwert, 1989; Bollerslev et al., 1992). In this section we generalize our

estimation method to take this into account. Because of our focus on maximum likelihood, a natural approach is to use the GARCH model of Bollerslev (1986). We will refer to this method as GARCH-MLE, and, for consistency, continue to refer to the method described in Section 2 as MLE. We ask three questions: (1) Do we still find a lower equity premium when we apply GARCH-MLE to the data? (2) Is GARCH-MLE efficient in small samples? (3) If we simulate data characterized by time-varying volatility and apply (homoskedastic, and therefore mis-specified) MLE, do we still find efficiency gains?

While the traditional GARCH model is typically applied to return data alone, our method closely relies on estimation of a bivariate process with correlated shocks. Allowing for time-varying volatility of returns but not of the dividend-price ratio seems artificial and unnecessarily restrictive. Following Bollerslev (1990), who estimates a GARCH model on exchange rates, we consider two correlated GARCH(1,1) processes. We assume

$$r_{t+1} - \mu_r = \beta(x_t - \mu_x) + u_{t+1} \tag{15a}$$

$$x_{t+1} - \mu_x = \theta(x_t - \mu_x) + v_{t+1}, \tag{15b}$$

where, conditional on information available up to and including time t,

$$\begin{bmatrix} u_{t+1} \\ v_{t+1} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2_{u,t+1} & \rho_{uv}\sigma_{u,t+1}\sigma_{v,t+1} \\ \rho_{uv}\sigma_{u,t+1}\sigma_{v,t+1} & \sigma^2_{v,t+1} \end{bmatrix}\right), \tag{15c}$$

with

$$\sigma^2_{u,t+1} = \omega_u + \alpha_u u_t^2 + \delta_u \sigma^2_{u,t}, \tag{15d}$$

$$\sigma^2_{v,t+1} = \omega_v + \alpha_v v_t^2 + \delta_v \sigma^2_{v,t}. \tag{15e}$$

We assume initial conditions

$$\sigma^2_{u,1} = \frac{\omega_u}{1 - \alpha_u - \delta_u},$$
$$\sigma^2_{v,1} = \frac{\omega_v}{1 - \alpha_v - \delta_v}.$$

Note that $\frac{\omega_u}{1-\alpha_u-\delta_u}$ and $\frac{\omega_v}{1-\alpha_v-\delta_v}$ represent the unconditional means of $\sigma^2_{u,t}$ and $\sigma^2_{v,t}$ respectively.[24] The bivariate GARCH(1,1) log-likelihood function is therefore

$$l(r_1,\ldots,r_T;x_1,\ldots,x_T|\mu_r,\mu_x,\beta,\theta,\omega_u,\alpha_u,\delta_u,\alpha_v,\delta_v,\rho_{uv},x_0) = \sum_{t=1}^{T}\log\left[(1-\rho^2_{uv})\sigma^2_{u,t}\sigma^2_{v,t}\right] + \frac{1}{1-\rho^2_{uv}}\sum_{t=2}^{T}\left(\frac{u_t^2}{\sigma^2_{u,t}} + 2\rho_{uv}\frac{u_t v_t}{\sqrt{\sigma^2_{u,t}\sigma^2_{v,t}}} + \frac{v_t^2}{\sigma^2_{v,t}}\right). \tag{16}$$

This likelihood function conditions on x_0, and thus is the GARCH analogue of the conditional maximum likelihood function (5). However, unlike in the homoskedastic case, there is no analytical expression for the unconditional distribution of x_0 (Diebold and Schuermann, 2000).[25] For this reason, we adopt a two-stage method that allows us both to estimate conditional heteroskedasticity, and to take into account the initial observation on the dividend-price ratio. While this represents a departure from "pure" maximum likelihood, it nonetheless allows us to consistently and efficiently estimate parameters.

We proceed as follows. First, we maximize the function (16) across the full

[24] Applying the law of iterated expectations, we find $Eu_t^2 = E[E_{t-1}u_t^2] = E\sigma^2_{u,t}$. The result for σ_u follows under stationarity by taking the expectation of the left and right hand sides of (15d), and the same argument works for σ_v.

[25] In principle we could capture this distribution by simulating from the conditional bivariate GARCH(1,1) over a long-period of time. To integrate this method into our optimization would not be easy however; for each function evaluation in our numerical optimization, we would need to simulate this distribution with enough accuracy to capture subtle effects of, say, the autoregressive coefficient θ along with the GARCH parameters. This would be challenging given that the parameter range of interest implies that x_t is highly persistent. We would then need to repeat the procedure thousands of times in our Monte Carlo simulations. It is hard to see the benefits (in terms of finite-sample efficiency gains) that this procedure would have over the more computationally feasible procedure that we do adopt.

set of parameters. We then maximize

$$
\begin{aligned}
&l(r_1,\ldots,r_T;x_0,\ldots,x_T|\mu_r,\mu_x,\beta,\theta,\omega_u,\alpha_u,\delta_u,\alpha_v,\delta_v,\rho_{uv}) = \\
&\qquad \log\left(\frac{\omega_v}{(1-\alpha_v-\delta_v)\left(1-\theta^2\right)}\right)+\frac{(x_0-\mu_x)^2}{\omega_v}(1-\alpha_v-\delta_v)\left(1-\theta^2\right) \\
&+\sum_{t=1}^{T}\log\left[(1-\rho_{uv}^2)\sigma_{u,t}^2\sigma_{v,t}^2\right]+\frac{1}{1-\rho_{uv}^2}\sum_{t=1}^{T}\left(\frac{u_t^2}{\sigma_{u,t}^2}+2\rho_{uv}\frac{u_tv_t}{\sqrt{\sigma_{u,t}^2\sigma_{v,t}^2}}+\frac{v_t^2}{\sigma_{v,t}^2}\right),
\end{aligned}
\tag{17}
$$

where we fix the estimates of ω_u, α_u, δ_u, ω_v, α_v, δ_v and ρ_{uv} from the first stage, and obtain new estimates of μ_r, μ_x, β and θ. The first two terms on the right hand side of (17) represents a density for the initial observation x_0. This density, which is normal with standard deviation $E[\sigma_{v,t}]/(1-\theta^2)$, represents an approximation to the true unknown density. By performing the estimation in two stages, we can make sure that the mis-specification in the second stage doesn't contaminate our GARCH estimation. Indeed, the GARCH estimation we perform in the first stage is the standard one in the literature. As mentioned above, we refer to this procedure as GARCH-MLE.

We report estimates in Table D.7. Similarly to previous studies (e.g. French et al. (1987)), we find that return volatility is moderately persistent, with a monthly autocorrelation of 0.72. Volatility of the dividend-price ratio is somewhat more persistent, with a monthly autocorrelation of 0.89. The average conditional volatilities of u_t and v_t are nearly identical to the unconditional volatilities in our benchmark case. Most importantly, given the focus of this study, the average equity premium is very close to what we found in our benchmark estimation: 0.335% per month, as opposed to 0.322%. The sample mean is 0.433% per month. Thus the finding of a lower equity premium is robust to time-varying volatility, which answers the first question we pose in the introduction to this section.

We now move on to the question of efficiency. We simulate 10,000 samples

from the process (15) using parameter values estimated by GARCH-MLE. We consider the performance of OLS (where we report sample means for the equity premium and the dividend-price ratio), the benchmark MLE procedure, and GARCH-MLE. Table 8 reports the means, standard deviations, and the 5th, 50th, and 95th percentiles of each parameter estimate.[26] We find that both MLE and GARCH-MLE are more efficient than the sample mean, and they are both about as efficient as each other. The efficiency gains are similar to what we see when the data generating process is homoskedastic (Table 6). We conclude that our estimation works well in the presence of time-varying volatility, both when we consider a method that explicitly takes time-varying volatility into account, and when we consider a (mis-specified) method that does not.

5.2. Non-stationarities in the dividend-price ratio

The previous section shows that our method works equally well for a bivariate GARCH(1,1) model as for our benchmark homoskedastic model. This may be because our method essentially translates information from long-run changes in the dividend-price ratio to information about returns. These long-run changes are sufficiently large that short-term volatility fluctuations do not alter their interpretations. Here, and in the sections that follow, we consider alternative models that have the potential to dramatically alter the interpretation of the time series of the dividend-price ratio, and thus the model's results for the equity premium. As in Section 4.2.2 where we set the correlation between shocks to the dividend-yield and returns to be zero, our aim is to "turn off" the gains from our method. However, in that case, a zero correlation was clearly counterfactual. Here, we consider models which, at least on a purely statistical

[26]For the volatility parameters σ_u and σ_v, we report the square root of the unconditional means of $\sigma^2_{u,t}$ and $\sigma^2_{v,t}$ for GARCH-MLE.

level, could account for the data. To focus on our main mechanism, we consider homoskedastic returns; however, the results of the previous section strongly suggest that these findings are also robust to conditional heteroskedasticity.

Given the observed high autocorrelation of the dividend-price ratio, a natural extension is to consider a random walk.[27] One immediate question that we face in assuming a random walk is the role of the predictive coefficient β. If the dividend-price ratio were to follow a random walk, and if β were nonzero, then the equity premium would be undefined. That is, excess stock returns, which would be non-stationary in this case, would not possess an unconditional mean. Any method, including the sample mean and our maximum likelihood procedure would give meaningless results. For this reason, when we consider a non-stationary dividend-price ratio (in this and in the subsequent section), we assume $\beta = 0$.

We therefore simulate 10,000 artificial samples from the process

$$r_{t+1} - \mu_r = u_{t+1}$$
$$x_{t+1} = x_t + v_{t+1}.$$

For each sample, we then apply our benchmark maximum likelihood procedure, as well as OLS regression.[28] For parameters μ_r and μ_x (this is a parameter in the estimation, not in the data generating process), we compare our maximum likelihood results with the sample means. Our benchmark maximum likelihood procedure (namely, maximizing Equation 6) is mis-specified because it assumes

[27] See for example, Campbell (2006) and Cochrane (2008).

[28] In our previous simulations, we initialize x_0 using a draw from the stationary distribution. Clearly this is not possible in this case. We report simulation results with x_0 set equal to its value in the data, but we have obtained identical results from randomizing over x_0. Other parameters are as follows: μ_r equals to its benchmark maximum likelihood estimate, σ_u the standard deviation of returns, σ_v the standard deviation of differences in the log dividend-price ratio, and ρ_{uv} to the correlation between returns and differences in the log dividend-price ratio.

stationarity and allows for predictability. Of course assumptions of OLS are also violated, as discussed above.

Table D.8 in the Online Appendix shows the results. Maximum likelihood still estimates the equity premium without bias, as shown by the fact that the average estimate of μ_r is exactly equal to the true value from the simulation. Besides correctly estimating the equity premium, maximum likelihood leads to significant gains in efficiency, even relative to our benchmark case. The standard deviation of the maximum likelihood estimate is only 30% of the standard deviation of the sample mean. The spread between the fifth and ninety-fifth percentile also falls by a factor greater than three. In this case, our estimation method does not pick up the non-stationarity in the dividend-price ratio (nor does OLS). However, the intuition of Section 4 still holds in this limiting case, and the model successfully estimates the equity premium with increased precision.

The previous discussion shows that our method is effective under a random-walk model for the dividend-price ratio. What about other forms of non-stationarity? Here, we consider what intuitively represents a worst-case scenario: a time trend in the dividend-price ratio. As in the case of the random walk model, we set β equal to zero so the equity premium is still well-defined. We therefore consider

$$r_{t+1} - \mu_r = u_{t+1} \tag{18a}$$

$$x_{t+1} - \mu_x = \Delta + \theta(x_t - \mu_x) + v_{t+1}, \tag{18b}$$

where Δ denotes the time trend. With the exceptions of Δ and β, we set the parameters to equal those of our benchmark calibration. We then set Δ so that the in-sample average of shocks to the dividend-price ratio is exactly zero. Because $\sum_{t=1}^{T} \hat{v}_t$ in the data is -1.051, and because the length of the sample is 707 months, this implies a value of Δ of -0.1487%.

We simulate 10,000 samples from (18). For each of these we compute OLS

and find the sample mean of the predictor variable and of the equity premium. We also run our benchmark maximum likelihood estimation, which is highly mis-specified in this case. For consistency, we continue to refer to this as maximum likelihood.

Results are shown in Table D.9 in the Online Appendix. Unlike in the case of the random walk, in this case mis-specification has serious consequences for the estimation of the equity premium. Whereas the sample mean finds, on average, the correct value, maximum likelihood finds a lower value: 0.280% versus 0.322%. To understand this result, consider that the true mean of x_t is undefined, but that in every sample there will be an average value of x_t. This average x_t will typically be lower than μ_x because the time trend makes x_t lower than it would be otherwise. The MLE for μ_x will be slightly higher than the sample mean because it will correct for what it sees as an unusual series of shocks (recall that we are still maximizing Equation 6). However, what appears to be an unusual series of shocks is in fact the time trend.

Now consider the estimation of the equity premium. Unlike the mean of x_t, the equity premium is well-defined because we have set β to equal zero. This is why the sample mean finds the correct answer. The maximum likelihood estimator, however, uses information from the predictor variable equation, information that is, in this case, incorrect. This information indicates that, on average, shocks have been positive to returns over each sample period, and thus it is necessary to adjust the equity premium downward.

While it would probably be nearly impossible to reject this time-trend model on purely statistical grounds, it seems unappealing from the point of view of economics. It implies that market participants would have known in advance about the decrease in the dividend-price ratio over the post-war sample, which is hard to believe. Not surprisingly given this basic intuition, equilibrium models of the asset prices tend to imply not (18), but rather the

autoregressive process (1b), at least as an approximation.[29]

5.3. Structural breaks

So far, we have assumed that a single process characterizes returns and the dividend-price ratio over the postwar period. Studies including Pástor and Stambaugh (2001), Lettau and Van Nieuwerburgh (2008) and Pettenuzzo and Timmermann (2011) argue that this period has been characterized by a structural break. The presence of a structural break could have several implications for our findings. Recall that the reason for our lower point estimate of the equity premium is the decline in the dividend-price ratio over the sample period. In a limiting case, where this decline is due entirely to a structural break, then our finding of a lower equity premium could completely disappear because the dividend-price ratio would no longer be declining over each sub-sample. As a related point, a structural break could make it less likely that we would find efficiency gains because, while the relevant sample size would be smaller, the persistence of the dividend-price ratio would be smaller as well.

To evaluate the effects, we use the framework of Lettau and Van Nieuwerburgh (2008), whose model is most similar to the one we consider. Lettau and van Nieuwerburgh find evidence for a structural break in the dividend-price ratio in 1994. Accordingly, we re-estimate our model on each sub-period. The results are reported in Table 9. This table shows that maximum likelihood still leads to substantially lower point estimates as compared with the sample mean. Consider first the 1953–1994 subperiod. This subperiod is characterized by relatively high returns, as indicated by a sample mean of 0.439%, slightly higher than our full sample average. However, this period is characterized by a striking decline in the dividend-price ratio, a fact that is largely undiminished

[29]Hansen et al. (2008) also present an example where a time-trend model for valuation ratios creates problems for interpretation of statistical findings. They argue similarly that the time trend model is an implausible description on economic grounds.

by breaking the sample in 1994 (see Figure 2). Our model thus attributes the high observed equity premium to an unusual series of shocks rather than a high true mean. The point estimate for the equity premium, at 0.315%, is *lower* than the point estimate for the full sample.

For the second sub-period, from 1995-2011, observed returns were lower, leading to a sample mean of 0.411%. Again, the dividend-price ratio declined over this sub-sample, so the maximum likelihood estimate is lower than the sample mean, at 0.336%. Thus maximum likelihood continues to have a substantial effect on the equity premium estimate, despite the presence of a structural break.

We now turn to the question of efficiency. Panel A1 of Table 10 shows simulation results when the parameters and the length of each fictitious sample are set to match the 1953–1994 subsample. We still do find efficiency gains, but they are indeed smaller than in our benchmark case. The standard error on the equity premium falls from 0.086 for the sample mean to 0.062 for maximum likelihood (in comparison, for our benchmark case, the sample mean had a standard error of 0.089 and the maximum likelihood estimate had a standard error of 0.050). Panel A1 also reveals the extent of the bias in the predictive and autoregressive coefficients. The mean estimate of β is substantially higher than its true value, and the mean of θ is substantially lower. This bias was also apparent in our benchmark case discussed in Section 3.6, but it is more substantial because of the reduction in sample size. Motivated by these results, we also consider a bias-corrected simulation, where, as before, we choose the true values of the parameters so that the mean in simulation matches the observed point estimates. As Panel A2 shows, the efficiency gain from maximum likelihood is almost as large as for our benchmark simulation when we correct for bias. The reason is that θ is higher than in Panel A1 (though it is still below the full-sample estimate), and the sample size is lower.

We repeat this analysis for the 1995–2011 subsample, with results shown

in Panel B. Panel B1 shows the results without the bias correction. In this case, because the sample size is so short, we still see efficiency gains despite the relatively low value of the autocorrelation. We also attempt a bias correction in Panel B2. Our results indicate the difficulties of inference over short time periods in the presence of persistent regressors. Even if we set the predictive coefficient to zero and the autocorrelation to 0.999, we are unable to quite match the values in the data (though we come close). Under this calibration, a short sample, combined with a high degree of persistence implies that the standard errors for maximum likelihood are less than half as large as for the sample mean. In other words, our efficiency gains are larger than even in the full sample.

To summarize, because a structural break does not entirely explain the decline in the price-dividend ratio, our method still produces substantially lower estimates of the equity premium than the sample mean, even when we take a structural break into account. Moreover, our efficiency gains are the same or larger than in our benchmark case.

6. Conclusion

A large literature has grown up around the empirical quantity known as the equity premium, in part because of its significance for evaluating models in macro-finance (Mehra and Prescott (1985)) and in part because of its practical significance as indicated by discussions in popular classics on investing (e.g. Siegel (1994), Malkiel (2003)) and in undergraduate and masters' level textbooks.

Estimation of the equity premium is almost always accomplished by taking sample means. The implicit assumption is that the period in question contains a representative sample of returns. We show that it is possible to relax this assumption, and obtain a better estimate of the premium, by bringing

additional information to bear on the problem, specifically the information contained separately in prices and dividends.

We show that the time series behavior of prices, dividends and returns, suggests that shocks to returns have been unusually positive over the post-war period. Thus the sample average will overstate the equity premium. We show that this intuition can be formalized with the standard econometric technique of maximum likelihood. Applying maximum likelihood rather than taking the sample average leads to an economically significant reduction in the equity premium of 1.3 percentage points from 6.4% to 5.1%. Furthermore, Monte Carlo experiments and RMSE calculations demonstrate that our method reduces sampling error and more reliably captures the true equity premium. We show similar results in international data and in characteristic-sorted portfolios. In particular, applying our results to portfolios sorted on the basis of market equity causes the well-known size premium to disappear.

Our method differs from the sample mean in that we require assumptions on the data generating process for the dividend-price ratio. We have shown that our findings are robust to a wide range of variations in these assumptions. Specifically, it is not necessary for returns to be homoskedastic, or even for the dividend-price ratio to be stationary. We also show that our method works well in the presence of structural breaks. The main conclusion from our findings is that the generous risk compensation offered by equities over the postwar sample may in part be an artifact of that period, and may not be a reliable guide to what investors will experience going forward.

References

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge, MA.

Andrews, D. W. K., 1993. Exactly median-unbiased estimation of first order autoregressive/unit root models. Econometrica 61, 139–165.

Bansal, R., Yaron, A., 2004. Risks for the long-run: A potential resolution of asset pricing puzzles. Journal of Finance 59, 1481–1509.

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

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.

Blanchard, O. J., 1993. Movements in the equity premium. Brookings Papers on Economic Activity 1993, 75–138.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., 1990. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. The Review of Economics and Statistics 72, pp. 498–505.

Bollerslev, T., Chou, R. Y., Kroner, K. F., 1992. {ARCH} modeling in finance: A review of the theory and empirical evidence. Journal of Econometrics 52, 5 – 59.

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

Boudoukh, J., Michaely, R., Richardson, M., Roberts, M. R., 2007. On the importance of measuring payout yield: Implications for empirical asset pricing. Journal of Finance 62, 877–915.

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

Campbell, J. Y., 2003. Consumption-based asset pricing. In: Constantinides, G., Harris, M., Stulz, R. (eds.), *Handbook of the Economics of Finance, vol. 1b*, Elsevier Science, North-Holland, pp. 803–887.

Campbell, J. Y., 2006. Household finance. Journal of Finance 61, 1553 – 1604.

Campbell, J. Y., Cochrane, J. H., 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. Journal of Political Economy 107, 205–251.

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

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

Campbell, J. Y., Viceira, L. M., 1999. Consumption and portfolio decisions when expected returns are time-varying. Quarterly Journal of Economics 114, 433–495.

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

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

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

Constantinides, G. M., 2002. Rational asset prices. The Journal of Finance 57, 1567–1591.

Davidson, R., MacKinnon, J. G., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York, New York.

DeLong, J. B., Magin, K., 2009. The u.s. equity return premium: Past, present, and future. The Journal of Economic Perspectives 23, 193–208.

Diebold, F. X., Mariano, R. S., 2002. Comparing predictive accuracy. Journal of Business & Economic Statistics 20, 134–144.

Diebold, F. X., Schuermann, T., 2000. Exact maximum likelihood estimation of observation-driven econometric models. In: R.S. Mariano, M. W., Schuermann, T. (eds.), *Simulation-Based Inference in Econometrics: Methods and Applications*, Cambridge University Press, pp. 205–217.

Donaldson, R. G., Kamstra, M. J., Kramer, L. A., 2010. Estimating the equity premium. Journal of Financial and Quantitative Analysis 45, 813–846.

Elliott, G., 1999. Efficient tests for a unit root when the initial observation is drawn from its unconditional distribution. International Economic Review 40, 767–783.

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., French, K. R., 1989. Business conditions and expected returns on stocks and bonds. Journal of Financial Economics 25, 23–49.

Fama, E. F., French, K. R., 1992. The cross-section of expected returns. Journal of Finance 47, 427–465.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on bonds and stocks. Journal of Financial Economics 33, 3–56.

Fama, E. F., French, K. R., 2001. Disappearing dividends: Changing firm characteristics or lower propensity to pay? Journal of Financial Economics 60, 3–43.

Fama, E. F., French, K. R., 2002. The equity premium. The Journal of Finance 57, pp. 637–659.

Ferson, W. E., Sarkissian, S., Simin, T. T., 2003. Spurious regressions in financial economics? Journal of Finance 58, 1393–1413.

French, K. R., Schwert, G. W., Stambaugh, R. F., 1987. Expected stock returns and volatility. Journal of Financial Economics 19, 3–29.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

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

Hansen, L. P., Heaton, J. C., Li, N., 2008. Consumption strikes back? Measuring long run risk. Journal of Political Economy 116, 260–302.

Hayashi, F., 2000. Econometrics. Princeton University Press, Princeton, NJ.

Ibbotson, R. G., Chen, P., 2003. Long-run stock returns: Participating in the real economy. Financial Analysts Journal 59, 88–98.

Jansson, M., Moreira, M. J., 2006. Optimal inference in regression models with nearly integrated regressors. Econometrica 74, 681–714.

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

Kelly, B., Pruitt, S., 2013. Market expectations in the cross-section of present values. The Journal of Finance 68, 1721–1756.

Kocherlakota, N. R., 1996. The equity premium: It's still a puzzle. Journal of Economic Literature 34, 42–71.

Larrain, B., Yogo, M., 2008. Does firm value move too much to be justified by subsequent changes in cash flow? Journal of Financial Economics 87, 200–226.

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

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

Lynch, A. W., Wachter, J. A., 2013. Using samples of unequal length in generalized method of moments estimation. Journal of Financial and Quantitative Analysis 48, 277–307.

Malkiel, B. G., 2003. A random walk down Wall Street. W. W. Norton and Company, Inc., New York, NY.

Mehra, R., Prescott, E., 1985. The equity premium puzzle. Journal of Monetary Economics 15, 145–161.

Merton, R. C., 1980. On estimating the expected return on the market: An exploratory investigation. Journal of Financial Economics 8, 323–361.

Müller, U. K., Elliott, G., 2003. Tests for unit roots and the initial condition. Econometrica 71, 1269–1286.

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

Pástor, Ľ., Stambaugh, R. F., 2001. The equity premium and structural breaks. The Journal of Finance 56, 1207–1239.

Pastor, L., Stambaugh, R. F., 2009. Predictive systems: Living with imperfect predictors. Journal of Finance 64, 1583 – 1628.

Pettenuzzo, D., Timmermann, A., 2011. Predictability of stock returns and asset allocation under structural breaks. Journal of Econometrics 164, 60–78.

Poirier, D. 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.

Schwert, W. G., 1989. Why does stock market volatility change over time? Journal of Finance 44, 1115–1153.

Shiller, R., 2000. Irrational Exuberance. Princeton University Press, Princeton, NJ.

Shiller, R. J., 1981. Do stock prices move too much to be justified by subsequent changes in dividends? American Economic Review 71, 421–436.

Siegel, J. J., 1994. Stocks for the long run: a guide to selecting markets for long-term growth. Irwin, Burr Ridge, IL.

Siegel, J. J., 2005. Perspectives on the equity risk premium. Financial Analysts Journal 61, 61–73.

Singleton, K., 2006. Empirical dynamic asset pricing: Model specification and econometric assessment. Princeton University Press, Princeton, NJ.

Stambaugh, R. F., 1997. Analyzing investments whose histories differ in length. Journal of Financial Economics 45, 285–331.

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

Torous, W., Valkanov, R., Yan, S., 2004. On predicting stock returns with nearly integrated explanatory variables. Journal of Business 77, 937–966.

van Binsbergen, J. H., Koijen, R. S. J., 2010. Predictive regressions: A present-value approach. The Journal of Finance 65, 1439–1471.

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

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

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

Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association 57, 348–368.

Zellner, A., 1986. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In: Goel, P., Zellner, A. (eds.), *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, North-Holland, Amsterdam, The Netherlands, pp. 233–243.

Table 1. Sample, Maximum Likelihood, and OLS Estimates.

	January 1953 – December 2011				January 1927 – December 2011			
	OLS	Sample	MLE	MLE_0	OLS	Sample	MLE	MLE_0
μ_r		0.433	0.322	0.312		0.464	0.391	0.395
μ_x		−3.545	−3.504	−3.437		−3.374	−3.383	−3.397
β	0.828		0.686		0.623		0.650	
θ	0.992		0.993	0.999	0.992		0.991	0.998
σ_u	4.414		4.416	4.426	5.466		5.464	5.473
σ_v	0.046		0.046	0.046	0.057		0.057	0.057
ρ_{uv}	−0.961		−0.961	−0.958	−0.953		−0.953	−0.952
RMSE		4.573	4.562	4.563		4.668	4.663	4.664
p(ΔMSE)			0.044	0.063			0.010	0.009

Notes: Estimation of the system

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where r_t is the continuously-compounded CRSP return minus the 30-day Treasury Bill return and x_t is the log of the dividend-price ratio. Shocks u_t and v_t are mean zero and iid over time with standard deviations σ_u and σ_v and correlation ρ_{uv}. Data are monthly. Means and standard deviations of returns are in percentage terms. In the OLS columns, parameters are estimated by ordinary least squares, with σ_u, σ_v, and ρ_{uv} estimated from the residuals. In the Sample column, μ_r is the average excess return over the sample and μ_x is the average of the log dividend-price ratio. In the MLE columns parameters are estimated using maximum likelihood. In the MLE_0 columns, parameters are estimated using maximum likelihood with the restriction $\beta = 0$. RMSE denotes the square root of the mean-squared error (MSE) from monthly out-of-sample return forecasts. $p(\Delta\text{MSE})$ denotes the p-value for a test of whether the MSE from out-of-sample forecasts generated by MLE differs from that generated by the sample mean.

Table 2. Estimates of the mean for characteristic-sorted portfolios

	Method	Estimate of μ_r by quintile					Premium
		Panel A: portfolios sorted by size					
		Small	Q2	Q3	Q4	Big	SmB
	Sample	0.957	0.978	0.975	0.936	0.797	0.160
μ_r	MLE	0.730	0.767	0.764	0.775	0.702	0.028
	MLE_0	0.709	0.752	0.758	0.777	0.689	0.020
	Sample	6.428	6.038	5.520	5.183	4.333	
RMSE	MLE	6.420	6.030	5.510	5.175	4.320	
	MLE_0	6.422	6.033	5.514	5.177	4.323	
p(ΔMSE)		0.282	0.236	0.178	0.138	0.022	
		Panel B: portfolios sorted by book-to-market ratio					
		Low	Q2	Q3	Q4	High	HmL
	Sample	0.755	0.845	0.930	0.991	1.074	0.319
μ_r	MLE	0.683	0.740	0.836	0.910	0.986	0.303
	MLE_0	0.631	0.735	0.840	0.903	1.013	0.382
	Sample	4.943	4.642	4.458	4.429	5.090	
RMSE	MLE	4.929	4.638	4.450	4.424	5.080	
	MLE_0	4.935	4.640	4.451	4.426	5.086	
p(ΔMSE)		0.191	0.329	0.040	0.299	0.175	

Notes: Estimates of μ_r (the expected net return) on characteristic-sorted portfolios in monthly data from 1953–2011. Estimates are reported in monthly percentage terms. Sample denotes the sample average of net returns. MLE denotes the maximum likelihood estimate of μ_r using the system

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where r_t is the portfolio return and x_t is the dividend-price ratio on the corresponding portfolio. MLE_0 denotes maximum likelihood with β restricted to be zero. Under the Premium column, we report the difference in the mean between the first and fifth quintile. RMSE denotes the square root of the mean-squared error (MSE) from monthly out-of-sample return forecasts. $p(\Delta\text{MSE})$ denotes the p-value for a test of whether the MSE from out-of-sample forecasts generated by MLE differs from that generated by the sample mean.

Table 3. Estimates for international indices

	Sample	MLE	MLE_0
All	0.362	0.191	0.249
Asia	0.259	0.119	0.130
EU with UK	0.423	0.327	0.360
EU without UK	0.386	0.243	0.321
Scandinavia	0.569	0.340	0.459

Notes: Estimates of the risk premium μ_r (the expected return less the riskfree rate) on international indices in monthly data beginning in January of 1976 and ending in 2011. Returns are dollar-denominated, and the U.S. 30-day Treasury Bill return proxies for the riskfree rate. Estimates are reported in monthly percentage terms. Sample denotes the sample average of excess returns. MLE denotes maximum likelihood of μ_r using the system

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where r_t is the index return in excess of the Treasury Bill and x_t is the dividend-price ratio for the index. MLE_0 denotes maximum likelihood with β restricted to be zero.

Table 4. Estimates of the mean for country-level indices

	Sample	MLE	MLE_0
Australia (1976)	0.463	0.429	0.423
Austria (1988)	0.404	−0.014	0.272
Belgium (1976)	0.511	0.344	0.349
Canada (1978)	0.473	0.134	0.200
Denmark (1990)	0.390	0.405	0.417
Finland (1989)	0.353	0.160	0.387
France (1976)	0.415	0.189	0.284
Germany (1976)	0.363	0.387	0.405
Hong Kong (1976)	0.631	0.690	0.688
Ireland (1992)	0.230	0.124	0.159
Italy (1976)	0.213	−0.191	0.042
Japan (1976)	0.198	0.063	0.040
Netherlands (1976)	0.530	0.445	0.449
New Zealand (1989)	0.121	−0.117	−0.010
Norway (1976)	0.474	0.357	0.392
Singapore (1976)	0.385	0.309	0.313
Spain (1976)	0.279	0.328	0.345
Sweden (1976)	0.630	0.408	0.534
Switzerland (1976)	0.465	0.286	0.417
UK (1976)	0.495	0.433	0.430

Notes: Estimates of the risk premium μ_r (the expected return less the riskfree rate) on country-level indices in monthly data beginning on the date in parentheses and ending in 2011. Returns are dollar-denominated, and the U.S. 30-day Treasury Bill return proxies for the riskfree rate. Estimates are reported in monthly percentage terms. Sample denotes the sample average of excess returns. MLE denotes maximum likelihood of μ_r using the system

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where r_t is the country return in excess of the Treasury Bill and x_t denotes the dividend-price ratio for the country. MLE_0 denotes maximum likelihood with β restricted to be zero.

Table 5. Estimates using alternative valuation measures

	log B/M ratio, 1977 – 2011				log E/P ratio, 1953 – 2011			
	OLS	Sample	MLE	MLE_0	OLS	Sample	MLE	MLE_0
Panel A: CRSP return in excess of the risk-free rate								
μ_r		0.427	0.304	0.274		0.433	0.384	0.371
μ_x		−0.739	−0.629	−0.574		−2.866	−2.839	−2.824
β	0.600		0.492		0.588		0.624	
θ	0.992		0.994	0.997	0.996		0.995	0.998
σ_u	4.653		4.651	4.660	4.419		4.416	4.426
σ_v	0.045		0.045	0.045	0.036		0.036	0.036
ρ_{uv}	−0.902		−0.902	−0.900	−0.698		−0.698	−0.697
RMSE		4.728	4.723	4.723		4.573	4.570	4.565
p(ΔMSE)			0.371	0.365			0.299	0.030
Panel B: S&P500 capital gain in excess of the risk-free rate								
μ_r		0.166	0.011	0.000		0.160	0.089	0.086
μ_x		−0.739	−0.629	−0.614		−2.866	−2.839	−2.830
β	0.270		0.164		0.149		0.191	
θ	0.992		0.994	0.995	0.996		0.995	0.997
σ_u	4.495		4.493	4.496	3.608		3.605	3.608
σ_v	0.045		0.045	0.045	0.036		0.036	0.036
ρ_{uv}	−0.914		−0.914	−0.914	−0.994		−0.994	−0.994
RMSE		4.576	4.578	4.578		3.681	3.677	3.675
p(ΔMSE)			0.566	0.545			0.248	0.151

Notes: Estimation of the system

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where r_t is the continuously-compounded return minus the 30-day Treasury Bill return and x_t is the logarithm of the book-to-market ratio or the inverse CAPE ratio (10-years inflation-adjusted earnings dividend by inflation-adjusted price), both for the S&P 500. In Panel A, the return is on the CRSP value-weighted portfolio. In Panel B, the return corresponds to the log price change on the S&P 500. Data are monthly. Shocks u_t and v_t are mean zero and iid over time with standard deviations σ_u and σ_v and correlation ρ_{uv}. Means and standard deviations of returns are in percentage terms. In the OLS columns, parameters are estimated by ordinary least squares, with σ_u, σ_v, and ρ_{uv} estimated from the residuals. In the Sample column, μ_r is the average excess return over the sample and μ_x is the average of x_t. In the MLE columns, parameters are estimated using maximum likelihood. In the MLE_0 columns, parameters are estimated using maximum likelihood with the restriction $\beta = 0$.

Table 6. Small-sample distribution of estimated parameters

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
			Panel A: Parameters set to estimates in Table 1				
μ_r	0.322	Sample	0.322	0.089	0.175	0.322	0.467
		MLE	0.323	0.050	0.241	0.324	0.404
μ_x	−3.504	Sample	−3.508	0.231	−3.894	−3.507	−3.126
		MLE	−3.508	0.221	−3.875	−3.507	−3.145
β	0.686	OLS	1.284	0.699	0.420	1.145	2.639
		MLE	1.243	0.670	0.440	1.103	2.541
θ	0.993	OLS	0.987	0.007	0.973	0.988	0.996
		MLE	0.987	0.007	0.974	0.989	0.996
σ_u	4.416	OLS	4.408	0.119	4.213	4.408	4.603
		MLE	4.406	0.119	4.211	4.406	4.600
σ_v	0.046	OLS	0.046	0.001	0.044	0.046	0.048
		MLE	0.046	0.001	0.044	0.046	0.048
ρ_{uv}	−0.961	OLS	−0.961	0.003	−0.965	−0.961	−0.956
		MLE	−0.961	0.003	−0.965	−0.961	−0.956
			Panel B: Bias Correction				
μ_r	0.322	Sample	0.324	0.138	0.097	0.327	0.546
		MLE	0.322	0.072	0.205	0.323	0.441
μ_x	−3.504	Sample	−3.510	0.582	−4.464	−3.512	−2.567
		MLE	−3.510	0.557	−4.425	−3.506	−2.601
β	0.090	OLS	0.750	0.643	−0.009	0.610	1.989
		MLE	0.686	0.601	0.036	0.528	1.881
θ	0.998	OLS	0.991	0.007	0.978	0.992	0.999
		MLE	0.992	0.006	0.979	0.993	0.998
σ_u	4.424	OLS	4.417	0.118	4.223	4.416	4.611
		MLE	4.417	0.118	4.225	4.416	4.612
σ_v	0.046	OLS	0.046	0.001	0.044	0.046	0.048
		MLE	0.046	0.001	0.044	0.046	0.048
ρ_{uv}	−0.961	OLS	−0.961	0.003	−0.965	−0.961	−0.956
		MLE	−0.961	0.003	−0.965	−0.961	−0.956

Notes: We simulate 10,000 monthly samples from the data generating process (DGP)

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where u_t and v_t are Gaussian and iid over time with standard deviations σ_u and σ_v and correlation ρ_{uv}. The sample length is as in postwar data. In Panel A parameters are set to their maximum likelihood estimates. In Panel B parameters are set to their maximum likelihood estimates with θ and β adjusted for bias. We conduct maximum likelihood estimation (MLE) for each sample path. As a comparison, we take sample means to estimate μ_r and μ_x (Sample) and use ordinary least squares to estimate the slope coefficients and the variance and correlations of the residuals (OLS). The table reports the means, standard deviations, and 5th, 50th, and 95th percentile values across simulations. The standard deviations correspond to small-sample standard errors for the postwar estimates in Table 1.

Table 7. Asymptotic standard errors for the 1953–2011 period

	Sample		MLE		MLE_0	
	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
μ_r	0.433	0.114	0.322	0.054	0.312	0.179
μ_x	−3.545	0.590	−3.504	0.279	−3.437	2.416
β			0.686	0.400		
θ			0.993	0.004	0.999	0.001
σ_u^2			19.498	0.223	19.587	0.237
σ_v^2			0.002	2.376×10^{-5}	0.002	2.521×10^{-5}
σ_{uv}			−0.194	6.446×10^{-7}	−0.193	7.179×10^{-7}

Notes: Point estimates and asymptotic standard errors for the system

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where r_t is the continuously-compounded CRSP return minus the 30-day Treasury Bill return and x_t is the log of the dividend-price ratio. Shocks u_t and v_t are mean zero and iid over time with variances σ_u^2 and σ_v^2 and covariance σ_{uv}. Data are monthly, January 1953 – December 2011. Returns are in percentage terms. In the Sample columns, parameters are estimated using the sample means. In the MLE columns, parameters are estimated using maximum likelihood. In the MLE_0 columns, parameters are estimated using maximum likelihood assuming $\beta = 0$.

Table 8. Small-sample distribution of estimators under conditional heteroskedasticity

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
		Sample	0.335	0.088	0.190	0.335	0.478
μ_r	0.335	MLE	0.335	0.049	0.253	0.335	0.415
		GARCH-MLE	0.335	0.049	0.252	0.335	0.414
		Sample	−3.570	0.225	−3.945	−3.570	−3.204
μ_x	−3.569	MLE	−3.571	0.214	−3.926	−3.572	−3.222
		GARCH-MLE	−3.571	0.214	−3.922	−3.571	−3.224
		OLS	1.288	0.694	0.425	1.156	2.621
β	0.689	MLE	1.244	0.668	0.436	1.103	2.554
		GARCH-MLE	1.236	0.664	0.436	1.100	2.531
		OLS	0.987	0.007	0.973	0.988	0.996
θ	0.993	MLE	0.987	0.007	0.974	0.989	0.996
		GARCH-MLE	0.987	0.007	0.974	0.989	0.996
		OLS	4.343	0.131	4.128	4.341	4.565
σ_u	4.351	MLE	4.342	0.131	4.126	4.340	4.563
		GARCH-MLE	4.341	0.133	4.125	4.339	4.566
		OLS	0.045	0.001	0.043	0.045	0.047
σ_v	0.045	MLE	0.045	0.001	0.043	0.045	0.047
		GARCH-MLE	0.045	0.001	0.043	0.045	0.047
		OLS	−0.959	0.003	−0.964	−0.959	−0.954
ρ_{uv}	−0.959	MLE	−0.959	0.003	−0.964	−0.959	−0.954
		GARCH-MLE	−0.959	0.003	−0.964	−0.960	−0.954

Notes: We simulate 10,000 monthly data samples from

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where u_t and v_t follow GARCH processes with conditional correlation ρ_{uv}. The parameter σ_u equals $\sqrt{E[\sigma_{ut}^2]}$ and similarly for σ_v. Parameters are set equal to estimates from GARCH-MLE as described in Section 5.1. For each sample path, we estimate parameters by OLS (and report sample means for μ_r and μ_x), by MLE (assuming homoskedastic shocks), and by GARCH-MLE.

Table 9. Sub-sample estimates

	January 1953 – December 1994				January 1995 – December 2011			
	OLS	Sample	MLE	MLE_0	OLS	Sample	MLE	MLE_0
μ_r		0.439	0.315	0.311		0.411	0.336	0.247
μ_x		−3.342	−3.337	−3.318		−4.048	−3.955	−3.845
β	2.538		2.186		2.614		1.968	
θ	0.977		0.981	0.999	0.972		0.979	0.995
σ_u	4.205		4.210	4.238	4.840		4.842	4.879
σ_v	0.043		0.043	0.043	0.051		0.051	0.051
ρ_{uv}	−0.967		−0.967	−0.960	−0.948		−0.949	−0.941
RMSE		4.413	4.398	4.399		5.129	5.150	5.121
p(ΔMSE)			0.014	0.033			0.823	0.329

Notes: Estimates of

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where u_t and v_t are Gaussian and iid over time with correlation ρ_{uv}. r_t is the continuously-compounded CRSP return minus the 30-day Treasury Bill return and x_t is the log of the dividend-price ratio. Two monthly data samples are considered: 1953–1994 and 1995–2011. Means and standard deviations of returns are in percentage terms. In the OLS columns, parameters are estimated by ordinary least squares, except for μ_r and μ_x, which are equal to the sample averages of excess returns and the log dividend-price ratio respectively. In the MLE columns, parameters are estimated using maximum likelihood. In the MLE_0 columns, parameters are estimated using g maximum likelihood under the restriction $\beta = 0$. RMSE denotes the square root of the mean-squared error (MSE) from monthly out-of-sample return forecasts. $p(\Delta\text{MSE})$ denotes the p-value for a test of whether the MSE from out-of-sample forecasts generated by MLE differs from that generated by the sample mean.

Table 10. Small-sample distribution of estimators in simulations calibrated to subsamples from Table 9

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95%
			Panel A1: 1953–1994, Calibration to MLE				
μ_r	0.315	Sample	0.315	0.086	0.176	0.315	0.457
		MLE	0.316	0.062	0.214	0.315	0.417
μ_x	−3.337	Sample	−3.336	0.097	−3.494	−3.337	−3.179
		MLE	−3.336	0.093	−3.488	−3.337	−3.183
β	2.186	MLE	2.983	1.133	1.518	2.776	5.122
θ	0.981	MLE	0.973	0.012	0.951	0.975	0.988
			Panel A2: 1953–1994, Bias Correction				
μ_r	0.315	Sample	0.315	0.115	0.125	0.314	0.504
		MLE	0.315	0.080	0.184	0.315	0.447
μ_x	−3.337	Sample	−3.336	0.166	−3.610	−3.337	−3.061
		MLE	−3.336	0.158	−3.595	−3.336	−3.074
β	1.400	MLE	2.185	0.961	1.007	1.983	4.066
θ	0.990	MLE	0.981	0.010	0.962	0.983	0.993
			Panel B1: 1995–2011, Calibration to MLE				
μ_r	0.336	Sample	0.333	0.187	0.028	0.332	0.639
		MLE	0.334	0.110	0.153	0.335	0.516
μ_x	−3.955	Sample	−3.952	0.145	−4.194	−3.951	−3.712
		MLE	−3.953	0.139	−4.183	−3.952	−3.721
β	1.968	MLE	3.841	2.220	1.158	3.358	8.071
θ	0.979	MLE	0.958	0.024	0.913	0.963	0.986
			Panel B2: 1995–2011, Bias Correction				
μ_r	0.336	Sample	0.331	0.339	−0.232	0.336	0.891
		MLE	0.332	0.152	0.083	0.332	0.582
μ_x	−3.955	Sample	−3.941	1.091	−5.741	−3.949	−2.161
		MLE	−3.941	1.079	−5.733	−3.952	−2.175
β	0	MLE	2.109	1.877	0.136	1.620	5.831
θ	0.999	MLE	0.976	0.020	0.937	0.981	0.996

Notes: We simulate 10,000 monthly samples from the data generating process (DGP)

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where u_t and v_t are Gaussian and iid over time with correlation ρ_{uv}. In Panel A, sample length and paramaters are for the 1953–1994 subsample, without bias correction (A1) and with bias correction (A2). In Panel B is constructed similarly for the 1995-2011 sample, except that here the bias-correction is partial. For each sample path, we conduct maximum likelihood estimation (MLE) and, for comparison, take sample means to find μ_r and μ_x (Sample). The table reports the means, standard deviations, and 5th, 50th, and 95th percentile values across simulations.

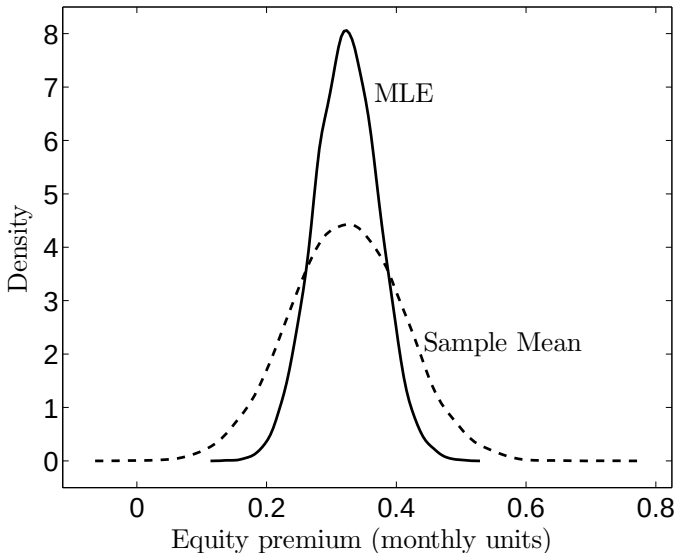


Fig. 1. Densities of the estimators of the equity premium in repeated samples of length equal to the postwar data. The solid line shows the density of the maximum likelihood estimate. The dashed line shows the density of the sample mean. Densities smoothed using a normal kernel.

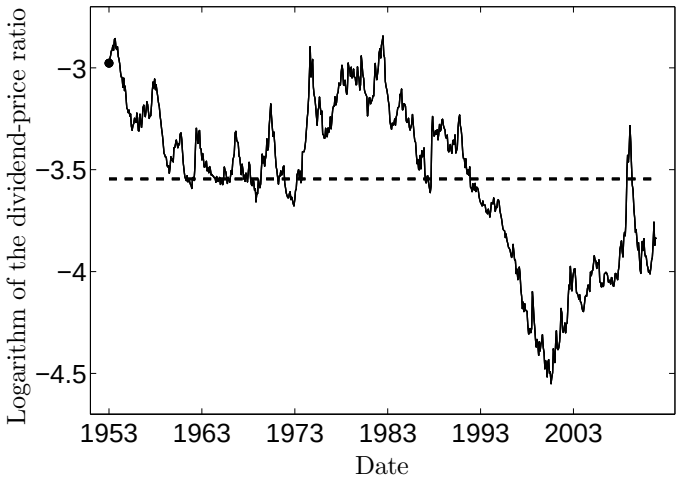


Fig. 2. The logarithm of the dividend-price ratio for the CRSP value-weighted portfolio. The dotted line indicates the mean, and the black dot the initial value.

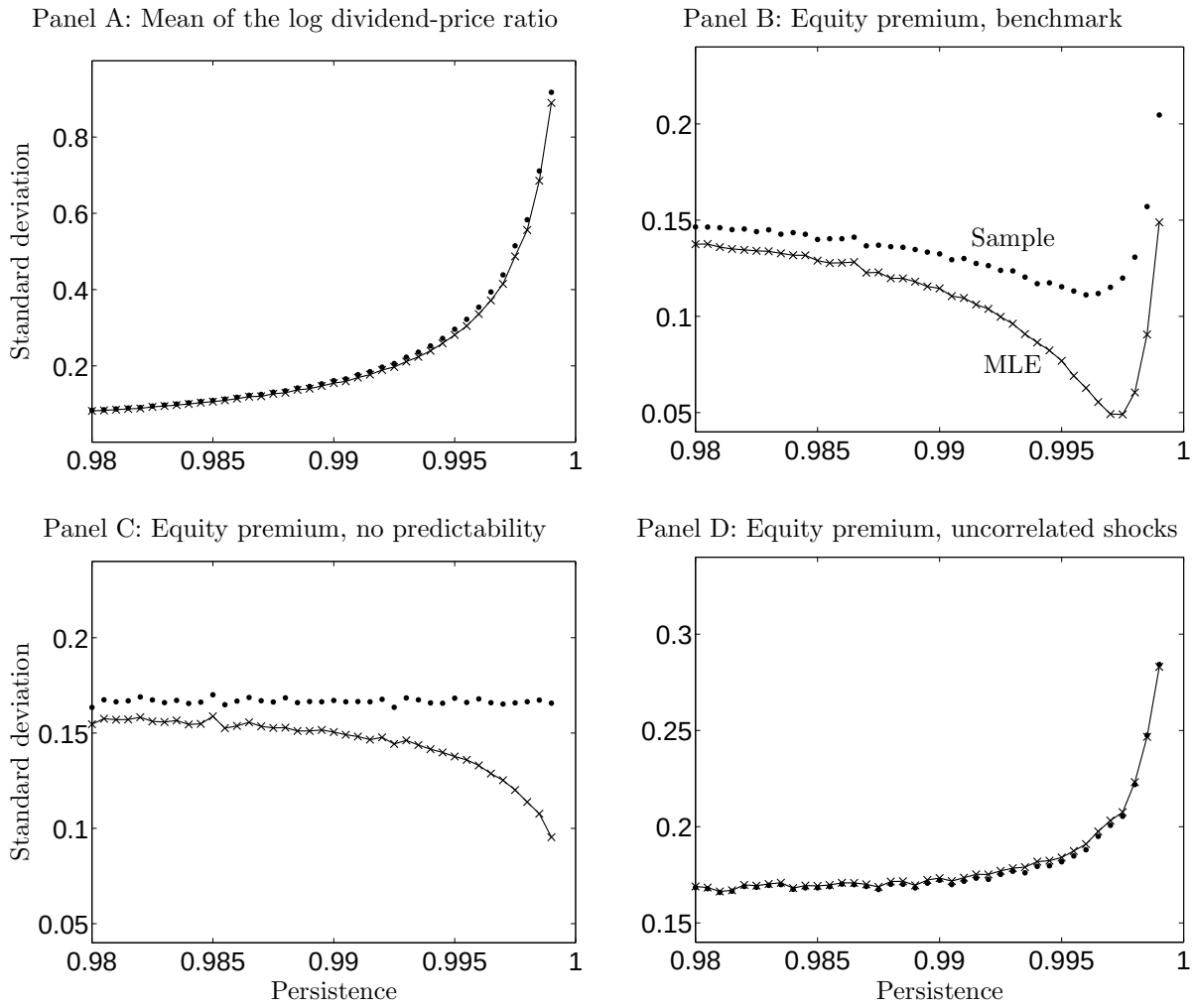


Fig. 3. Standard deviation of estimators of the mean of the log-dividend price ratio (Panel A) and of the equity premium (Panels B–D). Estimators are the sample mean (dots) and maximum likelihood (crosses). For each value of the autocorrelation θ, we simulate 10,000 monthly samples and calculate the standard deviation of estimates across samples. Parameters other than θ are set equal to their maximum likelihood estimates with the following exceptions. In Panel B, the predictive coefficient is bias-corrected. In Panel C, the predictive coefficient is set equal to zero. In Panel D, the predictive coefficient is bias-corrected and the correlation of the shocks is set equal to zero. Standard deviations on the mean return are in monthly percentage terms.

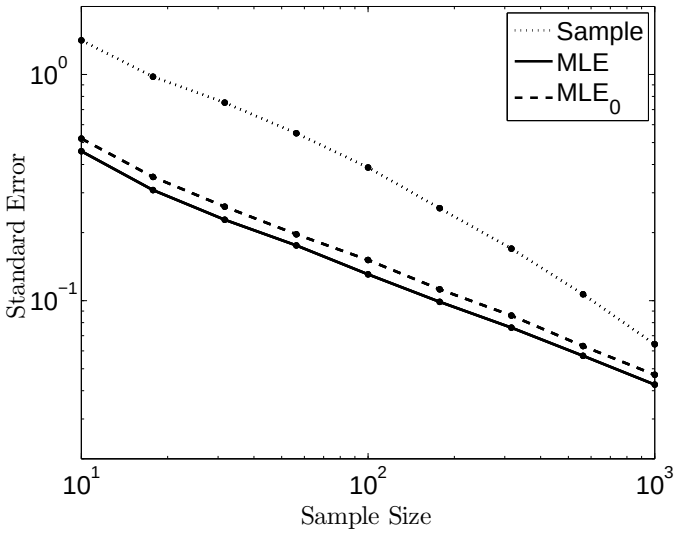
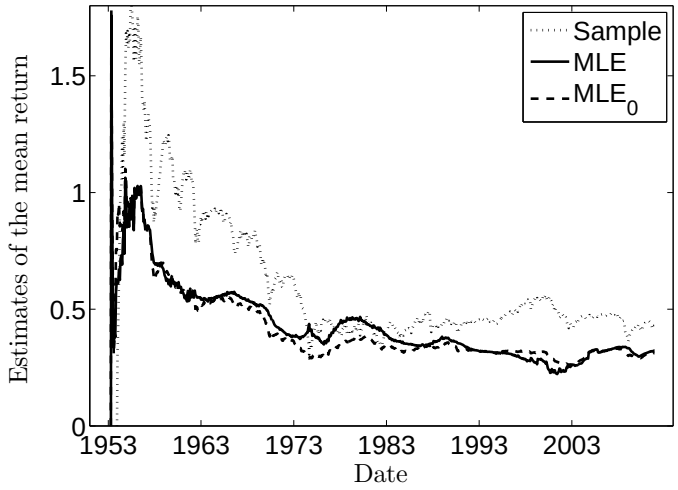


Fig. 4. Standard deviation of estimators of the equity premium as a function of sample size. For each T, we simulate 10,000 monthly samples of length T and calculate the standard deviation across samples. MLE_0 denotes maximum likelihood estimation with β restricted to be zero. Standard deviations are shown on a log-log scale. Standard deviations, which have the interpretation of standard errors on the estimates, are in monthly percentage terms.

Panel A: Estimates of the equity premium



Panel B: Difference in the estimates, scaled by $\sqrt{T}$

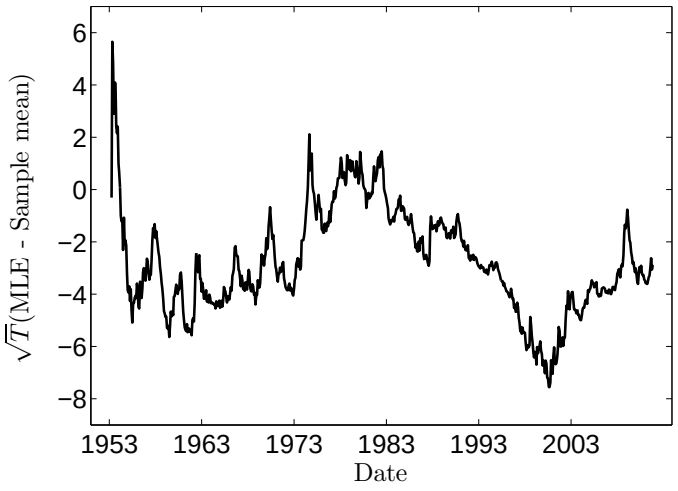


Fig. 5. For each month, beginning in January 1953, we estimate the equity premium using maximum likelihood (MLE), maximum likelihood with the $\beta = 0$ restriction (MLE_0), and the sample mean of returns less the riskfree rate (Sample) using data from January 1953 up until that month. The resulting time series is shown in Panel A. Panel B shows the difference between the maximum likelihood estimate and the sample mean, scaled by the length of the sample. Estimates of the equity premium are in monthly percentage terms.

Online Appendix

A. Derivation of the maximum likelihood estimators

A.1. Benchmark

We denote the maximum likelihood estimate of parameter q as $\hat{q}$. Here we derive the estimators for μ_r, μ_x, β, θ, σ_u^2, σ_v^2 and σ_{uv}. We note in particular that $\hat{\sigma}_u^2$ is the estimator of σ_u^2, not the square of the estimator of σ_u, and similarly for $\hat{\sigma}_v^2$. Maximizing the exact log likelihood function is the same as minimizing the function $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}(\beta,\theta,\mu_r,\mu_x,\sigma_{uv},\sigma_u,\sigma_v) = \log(\sigma_v^2) - \log(1-\theta^2) + \frac{1-\theta^2}{\sigma_v^2}(x_0-\mu_x)^2 \\ + T\log(|\Sigma|) + \frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T} u_t^2 - 2\frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T} u_t v_t + \frac{\sigma_u^2}{|\Sigma|}\sum_{t=1}^{T} v_t^2, \quad \text{(A.1)}\end{aligned}$$

where $|\Sigma| = \sigma_u^2\sigma_v^2 - \sigma_{uv}^2$. The function $\mathcal{L}$ is -2 times the logarithm of the likelihood function (6) modulo constants. The first-order conditions arise from setting the following partial derivatives of $\mathcal{L}$ to zero:

$$0 = \frac{\partial}{\partial\beta}\mathcal{L} = 2\left[\frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T} u_t(\mu_x - x_{t-1}) - \frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T}(\mu_x - x_{t-1})v_t\right] \quad \text{(A.2a)}$$

$$\begin{aligned}0 = \frac{\partial}{\partial\theta}\mathcal{L} = 2\Bigg[\frac{\theta}{1-\theta^2} - \theta\frac{(x_0-\mu_x)^2}{\sigma_v^2} \\ - \frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T} u_t(\mu_x - x_{t-1}) + \frac{\sigma_u^2}{|\Sigma|}\sum_{t=1}^{T} v_t(\mu_x - x_{t-1})\Bigg]\end{aligned} \quad \text{(A.2b)}$$

$$0 = \frac{\partial}{\partial\mu_r}\mathcal{L} = 2\left[-\frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T} u_t + \frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T} v_t\right] \quad \text{(A.2c)}$$

$$0 = \frac{\partial}{\partial \mu_x}\mathcal{L} = 2\Bigg[-\frac{1-\theta^2}{\sigma_v^2}(x_0 - \mu_x) + \frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T}\beta u_t \\ - \frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T}(\beta v_t - (1-\theta)u_t) - \frac{\sigma_u^2}{|\Sigma|}\sum_{t=1}^{T}(1-\theta)v_t\Bigg] \tag{A.2d}$$

$$0 = \frac{\partial}{\partial \sigma_{uv}}\mathcal{L} = -T\frac{2\sigma_{uv}}{|\Sigma|} \\ + 2\frac{\sigma_{uv}\sigma_v^2}{|\Sigma|^2}\sum_{t=1}^{T}u_t^2 - 2\frac{\sigma_u^2\sigma_v^2 + \sigma_{uv}^2}{|\Sigma|^2}\sum_{t=1}^{T}u_t v_t + 2\frac{\sigma_{uv}\sigma_u^2}{|\Sigma|^2}\sum_{t=1}^{T}v_t^2 \tag{A.2e}$$

$$0 = \frac{\partial}{\partial \sigma_u^2}\mathcal{L} = T\frac{\sigma_v^2}{|\Sigma|} - \frac{\sigma_v^4}{|\Sigma|^2}\sum_{t=1}^{T}u_t^2 + 2\frac{\sigma_{uv}\sigma_v^2}{|\Sigma|^2}\sum_{t=1}^{T}u_t v_t - \frac{\sigma_{uv}^2}{|\Sigma|^2}\sum_{t=1}^{T}v_t^2 \tag{A.2f}$$

$$0 = \frac{\partial}{\partial \sigma_v^2}\mathcal{L} = \frac{1}{\sigma_v^2} + T\frac{\sigma_u^2}{|\Sigma|} - (1-\theta^2)(x_0 - \mu_x)^2\frac{1}{\sigma_v^4} \\ - \frac{\sigma_{uv}^2}{|\Sigma|^2}\sum_{t=1}^{T}u_t^2 + 2\frac{\sigma_{uv}\sigma_u^2}{|\Sigma|^2}\sum_{t=1}^{T}u_t v_t - \frac{\sigma_u^4}{|\Sigma|^2}\sum_{t=1}^{T}v_t^2. \tag{A.2g}$$

Define the residuals

$$\hat{u}_t = r_t - \hat{\mu}_r - \hat{\beta}(x_{t-1} - \hat{\mu}_x) \tag{A.3a}$$

$$\hat{v}_t = x_t - \hat{\mu}_x - \hat{\theta}(x_{t-1} - \hat{\mu}_x). \tag{A.3b}$$

We now outline the algebra that allows us to solve these first-order conditions.

Step 1: Express $\hat{\mu}_x$ in terms of $\hat{\theta}$ and the data.

Combining the first-order conditions (A.2c) and (A.2d) gives

$$\sum_{t=1}^{T}\hat{v}_t = \left(1 + \hat{\theta}\right)(\hat{\mu}_x - x_0), \tag{A.4}$$

which we can write as

$$\hat{\mu}_x = \frac{\left(1+\hat{\theta}\right) x_0 + \sum_{t=1}^{T} \left(x_t - \hat{\theta} x_{t-1}\right)}{\left(1+\hat{\theta}\right) + \left(1-\hat{\theta}\right) T}. \tag{A.5}$$

Step 2: Express the covariance matrix in terms of $\hat{\mu}_x$, $\hat{\theta}$, $\hat{\mu}_r$, $\hat{\beta}$ and the data.

The first-order conditions (A.2e), (A.2f) and (A.2g) give the relations

$$T\hat{\sigma}_u^2 = -\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2}\hat{\sigma}_{uv} + (1-\hat{\theta}^2)(x_0 - \hat{\mu}_x)^2 \left(\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2}\right)^2 + \sum_{t=1}^{T} \hat{u}_t^2, \tag{A.6}$$

$$(T+1)\hat{\sigma}_v^2 = (1-\hat{\theta}^2)(x_0 - \hat{\mu}_x)^2 + \sum_{t=1}^{T} \hat{v}_t^2, \tag{A.7}$$

$$\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2} = \frac{\sum_{t=1}^{T} \hat{u}_t \hat{v}_t}{\sum_{t=1}^{T} \hat{v}_t^2}. \tag{A.8}$$

Step 3: Solve for $\hat{\theta}$ in terms of the data. This also gives $\hat{\mu}_x$ and $\hat{\sigma}_v^2$ in terms of the data.

Combining the first-order conditions (A.2a) and (A.2b) gives

$$0 = \sum_{t=1}^{T} (\hat{\mu}_x - x_{t-1})\hat{v}_t + \hat{\sigma}_v^2 \frac{\hat{\theta}}{1-\hat{\theta}^2} - \hat{\theta}(x_0 - \hat{\mu}_x)^2. \tag{A.9}$$

Here $\hat{\mu}_x$ and $\hat{v}_t$ are functions of only $\hat{\theta}$ and the data, so if we combine (A.27) and (A.7) we can get an equation for $\hat{\theta}$:

$$0 = (T+1) \sum_{t=1}^{T} (\hat{\mu}_x - x_{t-1})\hat{v}_t + \frac{\hat{\theta}}{1-\hat{\theta}^2} \sum_{t=1}^{T} \hat{v}_t^2 - T\hat{\theta}(x_0 - \hat{\mu}_x)^2. \tag{A.10}$$

Because we require that $-1 < \hat{\theta} < 1$, we can multiply this by

$$\left((T+1) - (T-1)\hat{\theta}\right)^2 \left(1-\hat{\theta}^2\right) \tag{A.11}$$

and rearrange to obtain

$$
\begin{aligned}
0 = T\left(\hat{\theta}-1\right)\left((T+1)\left(1-\hat{\theta}^2\right)+2\hat{\theta}\right)\left(\sum_{t=0}^{T} x_t - \hat{\theta}\sum_{t=1}^{T-1} x_t\right)^2 \\
+\left((T+1)-(T-1)\hat{\theta}\right)\left(\hat{\theta}-1\right)\left(\sum_{t=0}^{T} x_t - \hat{\theta}\sum_{t=1}^{T-1} x_t\right) \\
\times\left[2T\hat{\theta}(1+\hat{\theta})\left(\sum_{t=1}^{T-1} x_t\right)-\left((T+1)+(T-1)\hat{\theta}\right)\left(\sum_{t=0}^{T} x_t+\sum_{t=1}^{T-1} x_t\right)\right] \\
+\left((T+1)-(T-1)\hat{\theta}\right)^2 \\
\times\left[\hat{\theta}\left(\left(1-\hat{\theta}^2\right)T+1\right)\left(\sum_{t=1}^{T-1} x_t^2\right)+\left(\hat{\theta}^2(T-1)-(T+1)\right)\sum_{t=1}^{T} x_t x_{t-1}+\hat{\theta}\sum_{t=0}^{T} x_t^2\right].
\end{aligned}
\tag{A.12}
$$

This is a fifth-order polynomial in $\hat{\theta}$ where the coefficients are determined by the sample. As a consequence, it is very hard to establish analytical results on existence and uniqueness of solutions that would be accepted as estimators of θ. Nevertheless, in lengthy experimentation and simulation runs we have always found that this polynomial only has one root within the unit circle of the complex plane and that this root is real. Therefore this root is a valid MLE of θ. Given this solution for $\hat{\theta}$, (A.5) gives the estimator for μ_x and (A.7) gives the estimator for σ_v^2.

Step 4: Solve for $\hat{\mu}_r$ and $\hat{\beta}$ in terms of the data. This also gives the solution for $\hat{\sigma}_{uv}$ and $\hat{\sigma}_u^2$.

The first-order condition (A.2c) gives

$$
\sum_{t=1}^{T} \hat{u}_t = \frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2}\sum_{t=1}^{T} \hat{v}_t. \tag{A.13}
$$

Combining this with the first-order condition (A.2a) yields

$$
\hat{\beta} = \beta^{\mathrm{OLS}} + \frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2}\left(\hat{\theta}-\theta^{\mathrm{OLS}}\right), \tag{A.14}
$$

where

$$\theta^{\text{OLS}} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T} x_{t-1}^2 - \left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}\right)^2}\left[\frac{1}{T}\sum_{t=1}^{T} x_{t-1}x_t - \left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}\right)\left(\frac{1}{T}\sum_{s=1}^{T} x_s\right)\right] \tag{A.15}$$

is the OLS coefficient of regressing x_t on x_{t-1} and

$$\beta^{\text{OLS}} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T} x_{t-1}^2 - \left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}\right)^2}\left[\frac{1}{T}\sum_{t=1}^{T} x_{t-1}r_t - \left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}\right)\left(\frac{1}{T}\sum_{s=1}^{T} r_s\right)\right] \tag{A.16}$$

is the OLS coefficient of regressing r_t on x_{t-1}.

Equations (A.8), (A.13) and (A.14) constitute a system of three equations in the three unknowns $\hat{\mu}_r$, $\hat{\beta}$ and $\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2}$. The solution is

$$\begin{aligned}\hat{\mu}_r = \frac{1}{J}\Bigg[&\frac{1}{T}\sum_{t=1}^{T} r_t - \left(\frac{1}{T}\sum_{t=1}^{T} x_t - \hat{\mu}_x\right)\frac{F - \beta^{\text{OLS}}H}{1 + (\hat{\theta} - \theta^{\text{OLS}})H} \\ &- \left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1} - \hat{\mu}_x\right)\frac{\beta^{\text{OLS}}(1 + \hat{\theta}H) - \theta^{\text{OLS}}F}{1 + (\hat{\theta} - \theta^{\text{OLS}})H}\Bigg]\end{aligned} \tag{A.17}$$

$$\hat{\beta} = \frac{\beta^{\text{OLS}} + (\hat{\theta} - \theta^{\text{OLS}})F}{1 + (\hat{\theta} - \theta^{\text{OLS}})H} - \frac{(\hat{\theta} - \theta^{\text{OLS}})G}{1 + (\hat{\theta} - \theta^{\text{OLS}})H}\hat{\mu}_r \tag{A.18}$$

$$\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2} = \frac{F - \beta^{\text{OLS}}H}{1 + (\hat{\theta} - \theta^{\text{OLS}})H} - \frac{G}{1 + (\hat{\theta} - \theta^{\text{OLS}})H}\hat{\mu}_r, \tag{A.19}$$

where

$$J = 1 - \frac{G}{1 + (\hat{\theta} - \theta^{\text{OLS}})H}\left[\frac{1}{T}\sum_{t=1}^{T} x_t - \hat{\mu}_x - \theta^{\text{OLS}}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1} - \hat{\mu}_x\right)\right] \tag{A.20a}$$

$$F = \frac{\sum_{t=1}^{T} r_t\hat{v}_t}{\sum_{t=1}^{T} \hat{v}_t^2} \tag{A.20b}$$

$$G = \frac{\sum_{t=1}^{T} \hat{v}_t}{\sum_{t=1}^{T} \hat{v}_t^2} \tag{A.20c}$$

$$H = \frac{\sum_{t=1}^{T} (x_{t-1} - \hat{\mu}_x)\hat{v}_t}{\sum_{t=1}^{T} \hat{v}_t^2}. \tag{A.20d}$$

Expressions (A.17) and (A.18) provide the estimators for μ_r and β because they depend only on the data and $\hat{\mu}_x$ and $\hat{\theta}$, which we have already expressed in terms of the data. Finally, (A.19) gives the estimator the estimator of σ_{uv} via (A.7), which further yields the estimator of σ_u^2 via (A.6).

A.2. Restricted maximum likelihood

We consider maximum likelihood estimation under the restriction $\beta = 0$. We denote the restricted maximum likelihood estimate of parameter q as $\check{q}$. This case turns out to be less tractable than the unrestricted case, and for this reason, we fix the entries of the variance-covariance matrix Σ. We implement the estimator in two stages; in the first stage we run OLS to find Σ under the assumption of $\beta = 0$. In the second stage, we solve the equations that follow.

Consider (A.1) with the restriction of $\beta = 0$. The first-order conditions are as follows:

$$0 = \frac{\partial}{\partial \theta}\mathcal{L} = 2\Bigg[\frac{\theta}{1-\theta^2} - \theta\frac{(x_0-\mu_x)^2}{\sigma_v^2} - \frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T} u_t(\mu_x - x_{t-1}) + \frac{\sigma_u^2}{|\Sigma|}\sum_{t=1}^{T} v_t(\mu_x - x_{t-1})\Bigg] \tag{A.21a}$$

$$0 = \frac{\partial}{\partial \mu_r}\mathcal{L} = 2\left[-\frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T} u_t + \frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T} v_t\right] \tag{A.21b}$$

$$0 = \frac{\partial}{\partial \mu_x}\mathcal{L} = 2\Bigg[-\frac{1-\theta^2}{\sigma_v^2}(x_0-\mu_x) + \frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T}\beta u_t - \frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T}(\beta v_t - (1-\theta)u_t) - \frac{\sigma_u^2}{|\Sigma|}\sum_{t=1}^{T}(1-\theta)v_t\Bigg] \tag{A.21c}$$

Define the residuals

$$\check{u}_t = r_t - \check{\mu}_r \tag{A.22a}$$

$$\check{v}_t = x_t - \check{\mu}_x - \check{\theta}(x_{t-1} - \check{\mu}_x). \tag{A.22b}$$

We now outline the algebra that allows us to solve these first-order conditions.

Step 1: Express $\check{\mu}_x$ and $\check{\mu}_r$ in terms of $\check{\theta}$ and the data.

The first-order condition (A.21b) gives

$$\sum_{t=1}^{T} \check{u}_t = \frac{\sigma_{uv}}{\sigma_v^2} \sum_{t=1}^{T} \check{v}_t. \tag{A.23}$$

Combining this with the first-order condition (A.21c) gives

$$\sum_{t=1}^{T} \check{v}_t = \left(1 + \check{\theta}\right)\left(\check{\mu}_x - x_0\right), \tag{A.24}$$

which we can write as

$$\hat{\mu}_x = \frac{\left(1 + \check{\theta}\right) x_0 + \sum_{t=1}^{T} \left(x_t - \check{\theta} x_{t-1}\right)}{\left(1 + \check{\theta}\right) + \left(1 - \check{\theta}\right) T}. \tag{A.25}$$

Combining (A.24) and (A.23) yields

$$\check{\mu}_r = \frac{1}{T} \sum_{t=1}^{T} r_t - \frac{1}{T} \frac{\sigma_{uv}}{\sigma_v^2} \left(1 + \check{\theta}\right)\left(\check{\mu}_x - x_0\right). \tag{A.26}$$

Step 2: Solve for $\check{\theta}$ in terms of the data.

Substituting (A.23), (A.24) and (A.26) into the first-order condition (A.21a) gives

$$\begin{aligned} 0 = \sigma_v^2 \frac{\check{\theta}}{1 - \check{\theta}^2} - \check{\theta}(x_0 - \check{\mu}_x)^2 + \left(1 + \check{\theta}\right) \check{\mu}_x \left(\check{\mu}_x - x_0\right) \\ + \frac{1}{|\Sigma|} \left(\sum_{t=1}^{T} x_{t-1} \right) \left[\frac{\sigma_{uv}^2}{T} \left(1 + \check{\theta}\right)\left(\check{\mu}_x - x_0\right) + \sigma_u^2 \sigma_v^2 \left(1 - \check{\theta}\right) \check{\mu}_x \right] \\ + \frac{1}{|\Sigma|} \left[\sigma_{uv} \sigma_v^2 \sum_{t=1}^{T} x_{t-1} \left(r_t - \frac{1}{T} \sum_{s=1}^{T} r_s \right) - \sigma_u^2 \sigma_v^2 \sum_{t=1}^{T} x_{t-1} \left(x_t - \check{\theta} x_{t-1}\right) \right] \end{aligned} \tag{A.27}$$

Here $\breve{\mu}_x$ is a function of only $\breve{\theta}$ and the data, so given σ_u^2, σ_v^2 and σ_{uv} the above an equation for $\breve{\theta}$. Similarly to Appendix A, multiplying through by

$$\left((T+1)-(T-1)\breve{\theta}\right)^2\left(1-\breve{\theta}^2\right) \tag{A.28}$$

and carrying out the algebra gives a fifth-order polynomial in $\breve{\theta}$ where the coefficients are determined by the sample. As for the exact ML estimator in Appendix A, in lengthy experimentation and simulation runs we have always found that this polynomial only has one root within the unit circle of the complex plane and that this root is real. Therefore this root is a valid MLE of θ. Given this solution for $\breve{\theta}$, (A.25) gives the estimator for μ_x and (A.26) gives the estimator for μ_r.

A.3. The multivariate case

Our model is

$$\begin{aligned} r_{t+1}-\mu_r &= \sum_{i=1}^{N}\beta_i(x_{it}-\mu_{xi})+u_{t+1} \\ x_{1t+1}-\mu_{x1} &= \theta_1(x_{1t}-\mu_{x1})+v_{1t+1} \\ &\vdots \\ x_{Nt+1}-\mu_{xN} &= \theta_N(x_{Nt}-\mu_{xN})+v_{Nt+1} \end{aligned} \tag{A.29}$$

where, with $v_t=(v_{1t},\ldots,v_{Nt})^\top$, the vector $(u_t,v_t^\top)^\top$ is Gaussian and iid over time with covariance matrix

$$\Sigma=\begin{bmatrix} \sigma_u^2 & \sigma_{uv}^\top \\ \sigma_{uv} & \Sigma_v \end{bmatrix}. \tag{A.30}$$

Let Σ_x denote the covariance matrix of the vector $x_t=(x_{1t},\ldots,x_{Nt})^\top$. Element (i,j) of matrix Σ_x equals

$$\frac{\sigma_{ij}}{1-\theta_i\theta_j}, \tag{A.31}$$

where σ_{ij} is element (i, j) of matrix Σ_v. Let μ_x denote the vector $(\mu_{x1}, \ldots, \mu_{xN})^\top$, β denote the vector $(\beta_1, \ldots, \beta_N)^\top$, θ denote the vector $(\theta_1, \ldots, \theta_N)^\top$, and Θ denote the $N \times N$ diagonal matrix with the vector θ as its diagonal.

We denote the maximum likelihood estimate of parameter q as $\breve{q}$. Here we derive the estimators for μ_r, μ_x, β, and θ, taking σ_u^2, Σ_v, and σ_{uv} as given. Maximizing the exact log likelihood function is the same as minimizing the function $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}(\beta, \theta, \mu_r, \mu_x) = \log|\Sigma_x| + (x_0 - \mu_x)^\top \Sigma_x^{-1}(x_0 - \mu_x) \\ + T\log(|\Sigma|) + \sum_{t=1}^{T} \begin{pmatrix} u_t & v_t^\top \end{pmatrix} \Sigma^{-1} \begin{pmatrix} u_t \\ v_t \end{pmatrix}\end{aligned} \tag{A.32}$$

where $|Q|$ is notation for the determinant of matrix Q.

Let e_i denote a column vector with one as its ith element and zeros everywhere else. The first-order conditions arise from setting the partial derivatives of the likelihood function to zero.

$$0 = \frac{\partial}{\partial \beta_i}\mathcal{L} \Rightarrow 0 = \frac{\sigma_v^2}{|\Sigma|} \sum_{t=1}^{T} (\mu_x - x_{it-1}) \left(u_t - \sigma_{uv}^\top \Sigma_v^{-1} v_t\right) \tag{A.33a}$$

$$\begin{aligned}0 = \frac{\partial}{\partial \theta_i}\mathcal{L} \Rightarrow 0 = \operatorname{tr}\left(\Sigma_x^{-1}\frac{\partial}{\partial \theta_i}\Sigma_x\right) - (x_0 - \mu_x)^\top \Sigma_x^{-1}\left(\frac{\partial}{\partial \theta_i}\Sigma_x\right)\Sigma_x^{-1}(x_0 - \mu_x) \\ + 2\sum_{t=1}^{T}(x_{it-1} - \mu_{xi}) e_i^\top \left[\frac{1}{\sigma_\varepsilon^2}\Sigma_v^{-1}\sigma_{uv}u_t \right. \\ \left. - \left(\Sigma_v^{-1} + \frac{1}{\sigma_\varepsilon^2}\Sigma_v^{-1}\sigma_{uv}\sigma_{uv}^\top \Sigma_v^{-1}\right) v_t \right]\end{aligned} \tag{A.33b}$$

$$0 = \frac{\partial}{\partial \mu_r}\mathcal{L} \Rightarrow \sum_{t=1}^{T} u_t = \sigma_{uv}^\top \Sigma_v^{-1} \sum_{t=1}^{T} v_t \tag{A.33c}$$

$$0 = \frac{\partial}{\partial \mu_{xi}}\mathcal{L} \Rightarrow e_i^\top \Sigma_x^{-1}(x_0 - \mu_x) = (\theta_i - 1) \begin{pmatrix} 0 & e_i^\top \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \sum_{t=1}^{T} u_t \\ \sum_{t=1}^{T} v_t \end{pmatrix}, \tag{A.33d}$$

where

$$\sigma_\varepsilon^2 = \sigma_u^2 - \sigma_{uv}^\top \Sigma_v^{-1} \sigma_{uv}. \tag{A.34}$$

Define the residuals

$$\breve{u}_t = r_t - \breve{\mu}_r - \breve{\beta}^\top (x_{t-1} - \breve{\mu}_x) \tag{A.35a}$$

$$\breve{v}_t = x_t - \breve{\mu}_x - \breve{\Theta} (x_{t-1} - \breve{\mu}_x). \tag{A.35b}$$

We now outline the algebra that allows us to solve these first-order conditions.

Step 1: Express $\breve{\mu}_x$ in terms of $\breve{\Theta}$ and the data.

Stacking the first-order conditions for μ_{xi} in a vector, we get, after carrying out the algebra,

$$(\breve{\Theta} - \mathbb{I})\Sigma_v^{-1} \left[\sum_{t=1}^T \breve{v}_t + \frac{1}{\sigma_\varepsilon^2} \sigma_{uv} \left(\sigma_{uv}^\top \Sigma_v^{-1} \sum_{t=1}^T \breve{v}_t - \sum_{t=1}^T \breve{u}_t \right) \right] = \Sigma_v^{-1} (x_0 - \breve{\mu}_x). \tag{A.36}$$

Using (A.33c) we can simplify this to

$$\left(\breve{\Theta} - \mathbb{I} \right) \Sigma_v^{-1} \sum_{t=1}^T \breve{v}_t = \breve{\Sigma}_x^{-1} (x_0 - \breve{\mu}_x), \tag{A.37}$$

where $\breve{\Sigma}_x$ is a matrix with

$$\frac{\sigma_{ij}}{1 - \breve{\theta}_i \breve{\theta}_j} \tag{A.38}$$

as its (i,j)th element. We can write (A.37) as

$$\breve{\mu}_x = \left[\mathbb{I} + T \breve{\Sigma}_x \left(\breve{\Theta} - \mathbb{I} \right) \Sigma_v^{-1} \left(\breve{\Theta} - \mathbb{I} \right) \right]^{-1} \times \left[x_0 - \breve{\Sigma}_x \left(\breve{\Theta} - \mathbb{I} \right) \Sigma_v^{-1} \left(\sum_{t=1}^T x_t - \breve{\Theta} \sum_{t=1}^T x_{t-1} \right) \right]. \tag{A.39}$$

Given σ_u^2, Σ_v, and σ_{uv}, this equation expresses $\breve{\mu}_x$ in terms of the data and $\breve{\Theta}$.

Step 2: Solve for $\breve{\theta}$ in terms of the data. This also gives $\breve{\mu}_x$ in terms of the data.

Using (A.33a) in (A.33b) gives

$$0 = \operatorname{tr}\left(\breve{\Sigma}_x^{-1}\frac{\partial}{\partial\theta_i}\breve{\Sigma}_x\right) - (x_0-\mu_x)^\top\breve{\Sigma}_x^{-1}\left(\frac{\partial}{\partial\theta_i}\breve{\Sigma}_x\right)\breve{\Sigma}_x^{-1}(x_0-\mu_x) - 2e_i^\top\Sigma_v^{-1}\sum_{t=1}^{T}(x_{it-1}-\mu_{xi})\breve{v}_t, \quad \text{(A.40)}$$

for $i = 1,\ldots,N$. From (A.39) we have $\breve{\mu}_x$ in terms of $\breve{\theta}$ and the data, so if we combine (A.39) and (A.40) we get a system of N nonlinear equations for $\breve{\theta}_1,\ldots,\breve{\theta}_N$. Given the solution of this system for $\breve{\theta}_1,\ldots,\breve{\theta}_N$, (A.39) gives the estimator for μ_x.

Step 3: Solve for $\breve{\mu}_r$ and $\breve{\beta}$ in terms of the data.

The first-order condition (A.33c) gives

$$\breve{\mu}_r = \frac{1}{T}\sum_{t=1}^{T} r_t - \sigma_{uv}^\top\Sigma_v^{-1}\frac{1}{T}\sum_{t=1}^{T}\breve{v}_t - \breve{\beta}^\top\left(\frac{1}{T}\sum_{t=1}^{T}x_{t-1} - \breve{\mu}_x\right). \quad \text{(A.41)}$$

Using this in (A.33a) and carrying out the algebra we get

$$\begin{aligned}\left[\frac{1}{T}\sum_{t=1}^{T}x_{it-1}r_t - \left(\frac{1}{T}\sum_{t=1}^{T}x_{it-1}\right)\left(\frac{1}{T}\sum_{t=1}^{T}r_t\right)\right] \\ - \breve{\beta}^\top\left[\frac{1}{T}\sum_{t=1}^{T}x_{it-1}x_{t-1} - \left(\frac{1}{T}\sum_{t=1}^{T}x_{it-1}\right)\left(\frac{1}{T}\sum_{t=1}^{T}x_{t-1}\right)\right] \\ = \sigma_{uv}^\top\Sigma_v^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}x_{it-1}x_t - \left(\frac{1}{T}\sum_{t=1}^{T}x_{it-1}\right)\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right)\right. \\ \left. - \breve{\Theta}\left[\frac{1}{T}\sum_{t=1}^{T}x_{it-1}x_{t-1} - \left(\frac{1}{T}\sum_{t=1}^{T}x_{it-1}\right)\left(\frac{1}{T}\sum_{t=1}^{T}x_{t-1}\right)\right]\right\}, \quad \text{(A.42)}\end{aligned}$$

for $i = 1,\ldots,N$. Recall that we have solved for $\breve{\Theta}$ in terms of the data, so (A.42) constitutes a system of linear equations in $\breve{\beta}_1,\ldots,\breve{\beta}_N$. Given the solution of this system for $\breve{\beta}$, (A.41) gives the estimator for μ_r.

A.4. Asymptotic standard errors

Here we derive asymptotic standard errors for our maximum likelihood estimates using the methodology described in Hayashi (2000). Let q denote the vector

$$(\mu_r, \mu_x, \beta, \theta, \sigma_u^2, \sigma_v^2, \sigma_{uv})^\top, \tag{A.43}$$

and let s_t denote the score vector for observation t. In addition, let

$$p\left(x_0|q\right) = \left(2\pi\sigma_x^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{x_0 - \mu_x}{\sigma_x}\right)^2\right\} \tag{A.44}$$

denote the likelihood of the initial draw x_0, and let

$$p\left(u_t, v_t|q\right) = \left|2\pi\Sigma\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{\sigma_v^2}{|\Sigma|}u_t^2 - 2\frac{\sigma_{uv}}{|\Sigma|}u_t v_t + \frac{\sigma_u^2}{|\Sigma|}v_t^2\right)\right\} \tag{A.45}$$

denote the likelihood of the shock vector $(u_t, v_t)^\top$. We specify our objective function as $1/T$ times our exact likelihood function,

$$\frac{1}{T}\log p\left(r_1, \ldots, r_T; x_0, \ldots, x_T|q\right) = \frac{1}{T}\sum_{t=1}^{T}\left[\log p\left(u_t, v_t|q\right) + \frac{1}{T}p\left(x_0|q\right)\right], \tag{A.46}$$

where the equality follows by independence of the shocks over t, and by writing $p\left(x_0|q\right) = \sum_{t=1}^{T}\frac{1}{T}p\left(x_0|q\right)$. The score s_t is

$$s_t = \frac{\partial}{\partial q}\left[\log p\left(u_t, v_t|q\right) + \frac{1}{T}p\left(x_0|q\right)\right]. \tag{A.47}$$

We can see that the exact score is the conditional score $\frac{\partial}{\partial q}\log p\left(u_t, v_t|q\right)$ plus the "correction" term $\frac{\partial}{\partial q}\frac{1}{T}p\left(x_0|q\right)$.

The usual approach of obtaining the asymptotic covariance matrix is to derive a "sandwich estimator." Hayashi (2000, section 7.3) shows that, under maximum likelihood, the sandwich estimator simplifies due to the information matrix equality. One particularly convenient estimator of the asymptotic covariance matrix is

$$\text{Avar}\left(\hat{q}\right) = \left[\frac{1}{T}\sum_{t=1}^{T} s_t \; s_t^\top\right]^{-1}. \tag{A.48}$$

Hayashi notes that this estimator often has better finite-sample performance than the more complicated sandwich estimator, due the ease with which it is computed. The standard errors for our parameter estimates are given by the square root of the diagonal elements of $\text{Avar}(\hat{q})$ divided by $\sqrt{T}$.

It is straightforward to adopt the method above for restricted MLE; we set $\beta = 0$ and we drop the element of the score corresponding to β.

B. Further properties of maximum likelihood

B.1. The equity premium in levels

In this section we discuss how to translate our results for log returns into levels. For simplicity, assume that the log returns $\log(1 + R_t)$ are normally distributed. Then

$$E[R_t] = E\left[e^{\log(1+R_t)}\right] - 1 = e^{E[\log(1+R_t)]+\frac{1}{2}\text{Var}(\log(1+R_t))} - 1. \tag{B.1}$$

Using the definition of the excess log return, $E[\log(1 + R_t)] = E[r_t]+E[\log(1+R_t^f)]$, so the above implies that

$$E[R_t - R_t^f] = e^{E[r_t]}e^{E\left[\log\left(1+R_t^f\right)\right]+\frac{1}{2}\text{Var}(\log(1+R_t))} - 1 - E[R_t^f]. \tag{B.2}$$

Our maximum likelihood method provides an estimate of $E[r_t]$ and all other quantities above can be easily calculated using sample moments. Taking the sample mean of the series $R_t - R_t^f$ for the period 1953-2011 yields a risk premium that is 0.530% per month, or 6.37% per annum. On the other hand, using the above calculation and our maximum likelihood estimate of the mean of r_t gives an estimate of $\mathbb{E}[R_t - R_t^f]$ of 0.422% per month, or 5.06% per annum.[30] Thus our estimate of the risk premium in return levels is 131 basis

[30]In the data, in monthly terms for the period 1953-2011, the sample mean of R_t is 0.918%, the sample mean of R_t^f is 0.387%, the sample mean of $\log(1 + R_t^f)$ is 0.386% and the variance of $\log(1 + R_t)$ is 0.194%.

lower than taking the sample average, in line with our results for log returns.

B.2. Comparison with Fama and French (2002)

Fama and French (2002) also propose an estimator that takes the time series of the dividend-price ratio into account in estimating the mean return. Noting the following return identity:

$$R_t = \frac{D_t}{P_{t-1}} + \frac{P_t - P_{t-1}}{P_{t-1}}, \tag{B.3}$$

and taking the expectation:

$$E[R_t] = E\left[\frac{D_t}{P_{t-1}}\right] + E\left[\frac{P_t - P_{t-1}}{P_{t-1}}\right], \tag{B.4}$$

they propose replacing the capital gain term $E[(P_t - P_{t-1})/P_{t-1}]$ with dividend growth $E[(D_t - D_{t-1})/D_{t-1}]$. They argue that, because prices and dividends are cointegrated, their mean growth rates should be the same. They find that the resulting expected return is less than half the sample average, namely 4.74% rather than 9.62%.

While their argument seems intuitive, a closer look reveals a problem. Let $X_t = D_t/P_t$, and let lower-case letters denote natural logs. Then

$$d_{t+1} - d_t = x_{t+1} - x_t + p_{t+1} - p_t. \tag{B.5}$$

Because X_t is stationary, $E[x_{t+1} - x_t] = 0$ and it is indeed the case that

$$E[d_{t+1} - d_t] = E[p_{t+1} - p_t]. \tag{B.6}$$

However, exponentiating (B.5) and subtracting 1 implies

$$\frac{D_{t+1} - D_t}{D_t} = \frac{X_{t+1}}{X_t}\frac{P_{t+1}}{P_t} - 1. \tag{B.7}$$

That is, stationarity of X_t implies (B.6), but not $E[(P_t - P_{t-1})/P_{t-1}] = E[(D_t - D_{t-1})/D_{t-1}]$. Namely it does not imply that the average level growth rates are equal.

For expected growth rates to be equal in levels, (B.7) shows that it must be the case that $E\left[\frac{X_{t+1}}{X_t}\frac{P_{t+1}}{P_t}\right] = E\left[\frac{P_{t+1}}{P_t}\right]$. It seems unlikely that there are general conditions under which this holds. Note that it follows from $E[\log(X_{t+1}/X_t)] = 0$ and Jensen's inequality that $E[X_{t+1}/X_t] > 1$.[31] This implies that the estimator proposed by Fama and French (2002) is inconsistent for the equity premium, and thus it is not necessary (or possible) to evaluate efficiency.

Nonetheless, our results show that assuming cointegration of prices and dividends can be very informative for estimation of the mean return.[32] Indeed, the intuition that we will develop in the next section is closely related to that conjectured by Fama and French (2002): The sample average of realized returns is "too high" because shocks to discount rates (proxied for by the dividend-price ratio) were negative on average over the sample period.

[31]Indeed, if we assume that growth rates of dividends and prices are log-normal, a necessary and sufficient condition for equality of expected (level) growth rates is that the variances of the log growth rates are equal:

$$\mathrm{Var}(d_{t+1} - d_t) = \mathrm{Var}(p_{t+1} - p_t). \tag{B.8}$$

To see this, note that (B.6), combined with log-normality, implies that

$$E\left[\frac{D_{t+1}}{D_t}\right] e^{-\frac{1}{2}\mathrm{Var}(d_{t+1}-d_t)} = E\left[\frac{P_{t+1}}{P_t}\right] e^{-\frac{1}{2}\mathrm{Var}(p_{t+1}-p_t)}. \tag{B.9}$$

If (B.8) holds, then the second terms on the right and left hand side cancel, yielding the result. This is a knife-edge result in which the variance of the log dividend-price ratio x_t and the covariance of x_t with log price changes cancel out. However, it is well-known that prices are more volatile than dividends (Shiller, 1981).

[32]This point is also made by Constantinides (2002), who suggests adjusting the mean return by the difference in the valuation ratio between the first and last observation. Constantinides derives conditions such that the resulting estimator has lower variance than the average return.

C. Properties of the time series of returns under the benchmark data generating process

C.1. Mean reversion in returns

Consider the effect of a series of shocks on excess returns (in this subsection, we will assume, for expositional reasons, that the mean excess return is zero):

$$\begin{aligned} r_t &= \beta x_{t-1} + u_t \\ r_{t+1} &= \beta\theta x_{t-1} + \beta v_t + u_{t+1} \\ r_{t+2} &= \beta\theta^2 x_{t-1} + \beta\theta v_t + \beta v_{t+1} + u_{t+2} \end{aligned} \tag{C.1}$$

and so on. Thus, for $k \geq 1$, the autocovariance of returns is given by

$$\text{Cov}\left(r_t, r_{t+k}\right) = \theta^k \beta^2 \text{Var}(x_t) + \theta^{k-1}\beta\sigma_{uv}, \tag{C.2}$$

where $\text{Var}(x_t) = \sigma_v^2/(1-\theta^2)$. An increase in θ increases the variance of the predictor variable. In the absence of covariance between the shocks u and v, this effect would increase the autocovariance of returns through the term $\theta^k\beta^2\text{Var}(x_t)$. However, because u and v are negatively correlated, the second term in (C.2), $\theta^{k-1}\beta\sigma_{uv}$ is also negative. We show below that this second term dominates the first for all positive values of θ up until a critical value, at which point the first comes to dominate.

Assume $\theta > 0$, $\beta > 0$ and $\sigma_{uv} < 0$, as we estimate the case to be in our data. Substituting in $\text{Var}(x_t) = \sigma_v^2/(1-\theta^2)$, multiplying by $(1-\theta^2) > 0$ and dividing through by $\theta^{k-1}\beta > 0$ shows that the autocovariance of returns is negative whenever

$$-\sigma_{uv}\theta^2 + \beta\sigma_v^2\theta + \sigma_{uv} < 0. \tag{C.3}$$

The left-hand side is a quadratic polynomial in θ with a positive leading coefficient. As a result, whenever this polynomial has two real roots in θ, the entire expression is negative if and only if θ lies in between those roots. Indeed, the

polynomial has two real roots because its discriminant equals $\beta^2\sigma_v^4 + 4\sigma_{uv}^2 > 0$. Let θ_1 be the smaller of the two roots and let θ_2 be the larger one, that is,

$$\theta_2 = \frac{-\beta\sigma_v^2 + \sqrt{\beta^2\sigma_v^4 + 4\sigma_{uv}^2}}{-2\sigma_{uv}}. \tag{C.4}$$

Under our assumptions it is straightforward to prove that $\theta_1 < -1$ and $-1 < \theta_2 < 1$, so the only possible change of sign of the return autocovariance happens at θ_2. In particular, $\mathrm{Cov}\,(r_t, r_{t+k}) < 0$ whenever $\theta < \theta_2$ and $\mathrm{Cov}\,(r_t, r_{t+k}) > 0$ whenever $\theta > \theta_2$.

C.2. The variance of the sample mean return

By definition

$$\frac{1}{T}\sum_{t=1}^{T} r_t = \mu_r + \beta\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1} - \mu_x\right) + \frac{1}{T}\sum_{t=1}^{T} u_t, \tag{C.5}$$

thus

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} r_t\right) = \beta^2\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}\right) + \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} u_t\right) + 2\beta\mathrm{Cov}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}, \frac{1}{T}\sum_{t=1}^{T} u_t\right). \tag{C.6}$$

The variance of the average predictor is available and it depends on θ. The variance of the average residual does not depend on θ. Finally, the covariance of the average predictor and the average predictor depends on θ and ρ_{uv}. It is not a trivial quantity because even though u_t is uncorrelated with x_{t-1}, it is correlated with x_t via v_t whenever $\rho_{uv} \neq 0$ and thus it is also correlated with $x_{t+1}, x_{t+2}, \ldots, x_{T-1}$ whenever $\theta \neq 0$. In particular,

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} u_t\right) = \sigma_u^2\frac{1}{T}, \tag{C.7}$$

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}\right) = \frac{\sigma_v^2}{1-\theta^2}\left[\frac{1}{T}\left(1+2\frac{\theta}{1-\theta}\right)+\frac{2}{T^2}\frac{\theta(\theta^T-1)}{(1-\theta)^2}\right], \tag{C.8}$$

$$\mathrm{Cov}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}, \frac{1}{T}\sum_{t=1}^{T} u_t\right) = \sigma_{uv}\left[\frac{1}{T}\frac{1}{1-\theta}+\frac{1}{T^2}\frac{\theta^T-1}{(1-\theta)^2}\right], \tag{C.9}$$

so that

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} r_t\right) = \frac{1}{T}\left(\sigma_u^2+2\beta\frac{\sigma_{uv}}{1-\theta}+\beta^2\frac{\sigma_v^2}{1-\theta^2}\right) - \frac{1}{T^2}2\beta\frac{1-\theta^T}{(1-\theta)^2}\left(\beta\theta\frac{\sigma_v^2}{1-\theta^2}+\sigma_{uv}\right). \tag{C.10}$$

It follows that

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} r_t\right) = \frac{1}{T}\left(\sigma_u^2+\beta^2\frac{\sigma_v^2}{1-\theta^2}+2\beta\frac{\sigma_{uv}}{1-\theta}\right)+O\left(\frac{1}{T^2}\right). \tag{C.11}$$

The term $\sigma_u^2+\beta^2\sigma_v^2/(1-\theta^2)$ measures the contribution of the return shocks and the predictor to the variability of the sample-mean return. The term $\beta\sigma_{uv}/(1-\theta)$ measures the contribution of the covariance of the return shocks and the predictor shocks to the variability of the sample-mean return. The former term increases as θ increases, which says that the sample-mean return is more variable because the predictor is more variable. At the same time, the latter term becomes more negative as θ increases, so that in fact the overall variability of the sample-mean return can decrease.

D. Omitted tables and figures

Table D.1. Small-sample distribution of estimators: t-distributed shocks

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
μ_r	0.322	Sample	0.323	0.138	0.098	0.320	0.552
		MLE	0.322	0.072	0.204	0.322	0.440
μ_x	−3.504	Sample	−3.504	0.578	−4.454	−3.498	−2.543
		MLE	−3.504	0.549	−4.404	−3.498	−2.589
β	0.090	OLS	0.746	0.634	−0.007	0.601	1.947
		MLE	0.683	0.594	0.040	0.533	1.836
θ	0.998	OLS	0.991	0.007	0.978	0.993	0.999
		MLE	0.992	0.006	0.980	0.993	0.998
σ_u	4.430	OLS	4.419	0.185	4.136	4.411	4.727
		MLE	4.419	0.185	4.136	4.410	4.727
σ_v	0.046	OLS	0.046	0.002	0.043	0.045	0.049
		MLE	0.046	0.002	0.043	0.045	0.049
ρ_{uv}	−0.961	OLS	−0.961	0.004	−0.967	−0.961	−0.954
		MLE	−0.961	0.004	−0.967	−0.961	−0.954

Notes: We simulate 10,000 monthly samples from

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where $[u_t, v_t]$ has a bivariate t-distribution. The sample length is as in postwar data. Parameters are set to their maximum likelihood estimates (assuming normally distributed shocks) where β and θ are adjusted for bias. We conduct benchmark maximum likelihood estimation (MLE) for each sample path (this assumes normality and is therefore mis-specified). As a comparison, we take sample means to estimate μ_r and μ_x (Sample) and use ordinary least squares to estimate the slope coefficients and the variance and correlations of the residuals (OLS). The table reports the means, standard deviations, and 5th, 50th, and 95th percentile values across simulations. We set the degrees of freedom for the t-distribution to 5.96. This matches the average kurtosis of the estimated residuals for returns and the dividend-price ratio, and takes into account that the kurtosis is downward biased.

Table D.2. Small-sample distribution of estimators: Calibration to OLS estimates and sample means

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
μ_r	0.433	Sample	0.432	0.082	0.297	0.431	0.565
		MLE	0.432	0.049	0.352	0.432	0.513
μ_x	−3.545	Sample	−3.550	0.192	−3.865	−3.551	−3.232
		MLE	−3.550	0.184	−3.854	−3.552	−3.242
β	0.828	OLS	1.414	0.715	0.512	1.276	2.801
		MLE	1.372	0.689	0.515	1.241	2.675
θ	0.992	OLS	0.986	0.007	0.971	0.987	0.995
		MLE	0.986	0.007	0.972	0.988	0.995
σ_u	4.414	OLS	4.410	0.118	4.215	4.410	4.603
		MLE	4.408	0.118	4.214	4.408	4.601
σ_v	0.046	OLS	0.046	0.001	0.044	0.046	0.048
		MLE	0.046	0.001	0.044	0.046	0.048
ρ_{uv}	−0.961	OLS	−0.961	0.003	−0.965	−0.961	−0.956
		MLE	−0.961	0.003	−0.965	−0.961	−0.956

Notes: We simulate 10,000 monthly samples from

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where u_t and v_t are Gaussian and iid over time with standard deviations σ_u and σ_v and correlation ρ_{uv}. The sample length is as in postwar data. Parameters μ_r and μ_x are set to their sample averages, and parameters β, θ and variances and correlations are set to their OLS estimates. We conduct maximum likelihood estimation (MLE) for each sample path. We also report sample averages for μ_r and μ_x (Sample) and OLS estimates for the remaining parameters.

Table D.3. Small-sample distribution of estimators: calibration to 1927–2011 sample

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
μ_r	0.391	Sample	0.390	0.080	0.258	0.389	0.522
		MLE	0.391	0.058	0.295	0.390	0.485
μ_x	−3.383	Sample	−3.383	0.196	−3.710	−3.385	−3.063
		MLE	−3.384	0.190	−3.701	−3.384	−3.074
β	0.650	OLS	1.039	0.547	0.336	0.941	2.063
		MLE	1.018	0.530	0.345	0.923	2.007
θ	0.991	OLS	0.987	0.006	0.976	0.988	0.995
		MLE	0.987	0.006	0.977	0.989	0.994
σ_u	5.464	OLS	5.460	0.119	5.265	5.459	5.655
		MLE	5.458	0.119	5.263	5.458	5.653
σ_v	0.057	OLS	0.057	0.001	0.055	0.057	0.059
		MLE	0.057	0.001	0.055	0.057	0.059
ρ_{uv}	−0.953	OLS	−0.953	0.003	−0.958	−0.953	−0.948
		MLE	−0.953	0.003	−0.958	−0.953	−0.948

Notes: We simulate 10,000 monthly samples from

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where u_t and v_t are Gaussian and iid over time with standard deviations σ_u and σ_v and correlation ρ_{uv}. The sample length is set to match the 1927–2011 sample, and parameters are set to their maximum likelihood estimates over this period. We conduct maximum likelihood estimation (MLE) for each sample path. As a comparison, we take sample means to estimate μ_r and μ_x (Sample) and use ordinary least squares to estimate the slope coefficients and the variance and correlations of the residuals (OLS). The table reports the means, standard deviations, and 5th, 50th, and 95th percentile values across simulations.

Table D.4. Small-sample distribution of MLE_0

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95%
		Sample	0.312	0.169	0.040	0.309	0.591
μ_r	0.312	MLE	0.312	0.090	0.164	0.312	0.458
		MLE_0	0.312	0.089	0.164	0.312	0.460
		Sample	−3.439	1.078	−5.226	−3.450	−1.675
μ_x	−3.437	MLE	−3.436	1.051	−5.172	−3.438	−1.713
		MLE_0	−3.436	1.044	−5.156	−3.435	−1.718
		OLS	0.678	0.601	−0.048	0.550	1.845
β	0	MLE	0.602	0.558	0.012	0.450	1.694
		MLE_0					
		OLS	0.9920	0.0063	0.9798	0.9933	0.9996
θ	0.9992	MLE	0.9928	0.0058	0.9812	0.9944	0.9988
		MLE_0	0.9982	0.0012	0.9959	0.9985	0.9995

Notes: We simulate 10,000 monthly data samples from

$$\begin{aligned} r_{t+1} - \mu_r &= u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}. \end{aligned}$$

where u_t and v_t are Gaussian and iid over time with correlation ρ_{uv}. The sample length is as in postwar data. The parameters are set to their restricted maximum likelihood estimates in Table 1. For each sample path, we compute sample averages for μ_r and μ_x (Sample), OLS estimates of β and θ (OLS), unrestricted maximum likelihood (MLE, mis-specified in this case), and restricted maximum likelihood (MLE_0, correctly specified).

Table D.5. Estimates using multiple predictors

	returns	d/p	dsfp	tmsp
		Panel A: ML estimates		
μ_r	0.338			
μ_{xi}		−3.493	0.903	−0.871
β_i		0.893	−0.524	−0.143
θ_i		0.994	0.969	0.972
RMSE	4.569			
		Panel B: Sample and OLS estimates		
μ_r	0.441			
μ_{xi}		−3.548	0.904	−0.871
β_i		1.239	−0.157	−0.480
θ_i		0.991	0.968	0.973
RMSE	4.581			
		Panel C: Covariance matrix		
σ	4.391	0.046	0.101	0.246
ρ_{ui}		−0.957	−0.058	−0.115
ρ_{1i}			0.067	0.133
ρ_{2i}				−0.130

Notes: Estimates of

$$\begin{aligned} r_{t+1} - \mu_r &= \sum_{i=1}^{N} \beta_i (x_{it} - \mu_{x_i}) + u_{t+1} \\ x_{1,t+1} - \mu_{x_1} &= \theta_1 (x_{1t} - \mu_{x1}) + v_{1,t+1} \\ &\vdots \\ x_{N,t+1} - \mu_{x_N} &= \theta_N (x_{Nt} - \mu_{xN}) + v_{N,t+1} \end{aligned}$$

where u_t and $v_{1t}, \ldots, v_{Nt}$ are Gaussian and iid over time with covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \rho_{u1}\sigma_u\sigma_1 & \cdots & \rho_{uN}\sigma_u\sigma_N \\ \rho_{u1}\sigma_u\sigma_1 & \sigma_1^2 & \cdots & \rho_{1N}\sigma_1\sigma_N \\ \vdots & \vdots & \ddots & \\ \rho_{uN}\sigma_u\sigma_N & \rho_{1N}\sigma_1\sigma_N & & \sigma_N^2 \end{bmatrix},$$

where r_t is the continuously-compounded CRSP return minus the 30-day Treasury Bill return, x_{1t} is the log dividend-price ratio, x_{2t} is the default spread, and x_{3t} is the term spread. Data are monthly, April 1953 – December 2011. Means and standard deviations of returns are in percentage terms. In Panel A, parameters are estimated using maximum likelihood. In Panel B, μ_r and μ_{x_i} are estimated by sample averages, and β_i and θ_i are estimated by ordinary least squares. Panel C gives the standard deviations of the shocks (top row) and the correlations between the shocks estimated using OLS residuals. Variables are the dividend-price ratio (d/p), the continuously-compounded yield of BAA-rated bonds minus the continuously-compounded yield of AAA rated bonds (dfsp), and the continuously-compounded yield of ten-year treasury bonds minus the continuously-compounded yield of one-year treasury bonds (tmsp).

Table D.6. Annual estimates using repurchase-adjusted dividend-price ratios

	Treasury-stock adjusted d/p				Cash-flow adjusted d/p			
	OLS	Sample	MLE	MLE_0	OLS	Sample	MLE	MLE_0
μ_r		5.718	4.252	4.092		5.718	4.806	4.558
μ_x		−3.352	−3.334	−3.318		−3.258	−3.240	−3.221
β	19.556		17.221		21.343		19.868	
θ	0.897		0.923	0.977	0.865		0.883	0.958
σ_u	16.164		16.185	17.195	16.167		16.113	17.195
σ_v	0.125		0.126	0.125	0.130		0.130	0.130
ρ_{uv}	−0.700		−0.708	−0.658	−0.668		−0.674	−0.628
RMSE		17.233	16.470	16.598		17.233	16.581	16.606
p(ΔMSE)			0.021	0.102			0.023	0.094

Notes: Estimates of

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where r_t is the continuously-compounded CRSP return minus the annual Treasury Bill return and x_t is the logarithm of the dividend yield, adjusted for repurchases. Two such adjusted dividend-price ratios are considered: the cash-flow based yield (cfby) and the Treasury-stock based yield (tsby). Shocks u_t and v_t are mean zero and iid over time with standard deviations σ_u and σ_v and correlation ρ_{uv}. Return data and dividend-yield data are annual, 1953–2003. Means and standard deviations of returns are in percentage terms. Under the OLS columns, parameters are estimated by ordinary least squares, with σ_u, σ_v, and ρ_{uv} estimated from the residuals. In the Sample column, μ_r is the average excess return over the sample and μ_x is the average of the log dividend-price ratio. In the MLE column parameters are estimated using maximum likelihood. In the MLE_0 columns, parameters are estimated using maximum likelihood assuming $\beta = 0$. RMSE denotes the root-mean-squared error from monthly out-of-sample return forecasts.

Table D.7. Estimation of a predictive regression with heteroskedasticity

Panel A: Means and coefficients		Panel B: Volatility parameters		Panel C: Covariance matrix	
μ_r	0.335	ω_u	4.763	σ_u^*	4.351
μ_x	−3.569	α_u	0.029	σ_v^*	0.045
β	0.688	δ_u	0.719	ρ_{uv}	−0.959
θ	0.993	ω_v	1.855×10^{-4}		
		α_v	0.016		
		δ_v	0.892		

Notes: We estimate the bivariate process

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}, \end{aligned}$$

where, conditional on information available up to and including time t,

$$\begin{bmatrix} u_{t+1} \\ v_{t+1} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{u,t+1}^2 & \rho_{uv}\sigma_{u,t+1}\sigma_{v,t+1} \\ \rho_{uv}\sigma_{u,t+1}\sigma_{v,t+1} & \sigma_{v,t+1}^2 \end{bmatrix}\right),$$

and

$$\begin{aligned} \sigma_{u,t+1}^2 &= \omega_u + \alpha_u u_t^2 + \delta_u \sigma_{u,t}^2, \\ \sigma_{v,t+1}^2 &= \omega_v + \alpha_v v_t^2 + \delta_v \sigma_{v,t}^2. \end{aligned}$$

Here, r_t is the continuously compounded return on the value-weighted CRSP portfolio in excess of the return on the 30-day Treasury Bill and x_t is the log of the dividend-price ratio. Starred parameters are implied by other estimates, namely $\sigma_u^* = \sqrt{\omega_u/(1-\alpha_u-\delta_u)}$ and $\sigma_v^* = \sqrt{\omega_v/(1-\alpha_v-\delta_v)}$. Parameters are estimated using a two-stage process by which the means and coefficients (Panel A) are treated as fixed and the volatility parameters (Panels B and C) are estimated using conditional maximum likelihood in the first stage, and the volatility parameters are treated as fixed, while the means and coefficients are re-estimated in the second stage. Data are monthly, from January 1953 to December 2011. Means and standard deviations of returns are in percentage terms.

Table D.8. Small-sample distribution of estimators when the dividend-price ratio follows a random walk

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
μ_r	0.322	Sample	0.325	0.166	0.050	0.327	0.599
		MLE	0.322	0.047	0.246	0.323	0.401
μ_x	−3.504	Sample	−2.988	0.699	−4.130	−2.996	−1.845
		MLE	−2.986	0.637	−4.006	−2.997	−1.971
θ	0.993	OLS	0.992	0.006	0.980	0.994	1.000
		MLE	0.993	0.006	0.981	0.995	0.999
σ_u	4.416	OLS	4.413	0.117	4.221	4.414	4.605
		MLE	4.415	0.117	4.223	4.417	4.607
σ_v	0.046	OLS	0.046	0.001	0.044	0.046	0.048
		MLE	0.046	0.001	0.044	0.046	0.048
ρ_{uv}	−0.961	OLS	−0.962	0.003	−0.967	−0.962	−0.957
		MLE	−0.962	0.003	−0.967	−0.962	−0.957

Notes: We simulate 10,000 monthly data samples from

$$\begin{aligned} r_{t+1} - \mu_r &= u_{t+1} \\ x_{t+1} &= x_t + v_{t+1} \end{aligned}$$

where u_t and v_t are Gaussian and iid over time with correlation ρ_{uv}. For each sample path we conduct (mis-specified) maximum likelihood estimation (MLE) of

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}. \end{aligned}$$

For comparison, we take sample means to estimate μ_r and μ_x (Sample) and use ordinary least squares to estimate the slope coefficients and the variance and correlations of the residuals (OLS). The table reports the means, standard deviations, and 5th, 50th, and 95th percentile values across simulations.

Table D.9. Small-sample distribution of estimators when the dividend-price ratio has a time trend

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
μ_r	0.322	Sample	0.322	0.168	0.044	0.321	0.599
		MLE	0.280	0.145	0.044	0.280	0.516
μ_x	−3.504	Sample	−3.682	0.234	−4.066	−3.682	−3.292
		MLE	−3.663	0.223	−4.028	−3.661	−3.296
β	0	OLS	0.590	0.684	−0.255	0.460	1.880
		MLE	0.514	0.660	−0.270	0.375	1.756
θ	0.993	OLS	0.987	0.007	0.974	0.988	0.996
		MLE	0.988	0.007	0.975	0.989	0.996
σ_u	4.416	OLS	4.410	0.117	4.219	4.410	4.602
		MLE	4.409	0.117	4.218	4.410	4.601
σ_v	0.046	OLS	0.046	0.001	0.044	0.046	0.048
		MLE	0.046	0.001	0.044	0.046	0.048
ρ_{uv}	−0.961	OLS	−0.961	0.003	−0.965	−0.961	−0.956
		MLE	−0.961	0.003	−0.965	−0.961	−0.956

Notes: We simulate 10,000 monthly data samples from

$$\begin{aligned} r_{t+1} - \mu_r &= u_{t+1} \\ x_{t+1} - \mu_x &= \Delta + \theta(x_t - \mu_x) + v_{t+1} \end{aligned}$$

where u_t and v_t are Gaussian and iid over time with correlation ρ_{uv}. We set μ_r, μ_x, θ, σ_u, σ_v and ρ_{uv} to their benchmark maximum likelihood estimates (Table 1) and Δ to the mean residual $(1/T)\sum_{t=1}^{T} \hat{v}_t = -0.14868$. For each sample path we conduct (mis-specified) maximum likelihood estimation (MLE) of

$$\begin{aligned} r_{t+1} - \mu_r &= \beta(x_t - \mu_x) + u_{t+1} \\ x_{t+1} - \mu_x &= \theta(x_t - \mu_x) + v_{t+1}. \end{aligned}$$

For comparison, we take sample means to estimate μ_r and μ_x (Sample) and use ordinary least squares to estimate the slope coefficients and the variance and correlations of the residuals (OLS). The table reports the means, standard deviations, and 5th, 50th, and 95th percentile values across simulations.

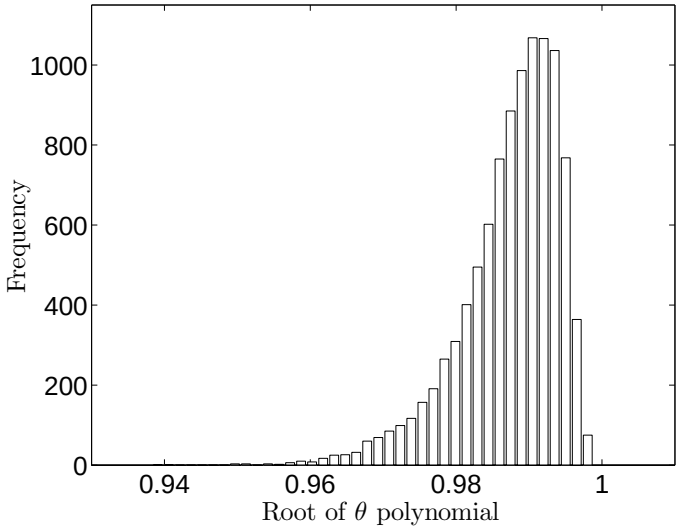


Fig. D.1. Histogram of maximum likelihood estimates of θ, the autocorrelation of the dividend-price ratio from simulated data. We simulate 10,000 monthly data samples from (1) with length and parameters as in the postwar data series.

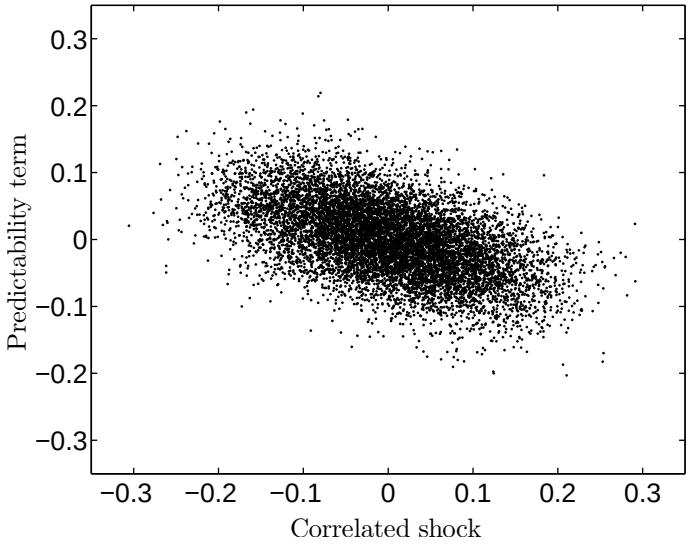


Fig. D.2. We simulate 10,000 monthly data samples from (1) with length and parameters as in the postwar data series. The figure shows the joint distribution of the predictability term $\hat{\beta}\frac{1}{T}\sum_{t=1}^{T}(x_{t-1}-\hat{\mu}_x)$ and the correlated shock term $\frac{1}{T}\sum_{t=1}^{T}\hat{u}_t$ that sum to the difference between the maximum likelihood estimate and the sample mean.

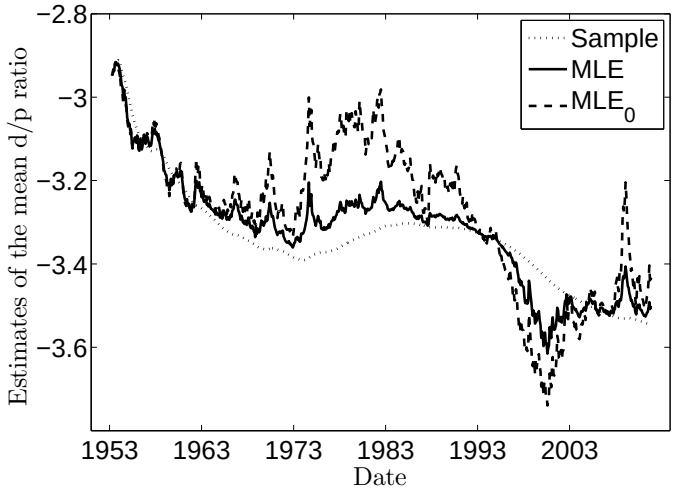


Fig. D.3. For each month, beginning in January 1953, we estimate the mean of the dividend-price ratio using maximum likelihood (MLE), maximum likelihood with the restriction $\beta = 0$ (MLE_0), and the sample mean (Sample), using data from January 1953 up until that month.

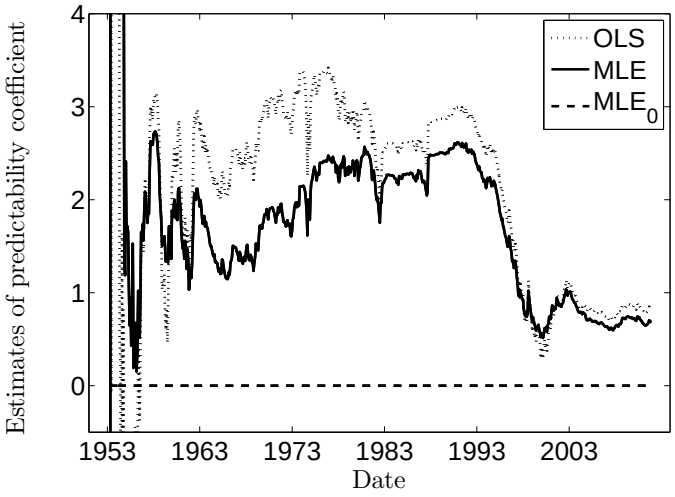


Fig. D.4. For each month, beginning in January 1953, we estimate the coefficient of predictability (β) using maximum likelihood (MLE), and Ordinary Least Squares (OLS), using data from January 1953 up until that month. For our restricted maximum likelihood method (MLE_0), $\beta = 0$ by assumption.

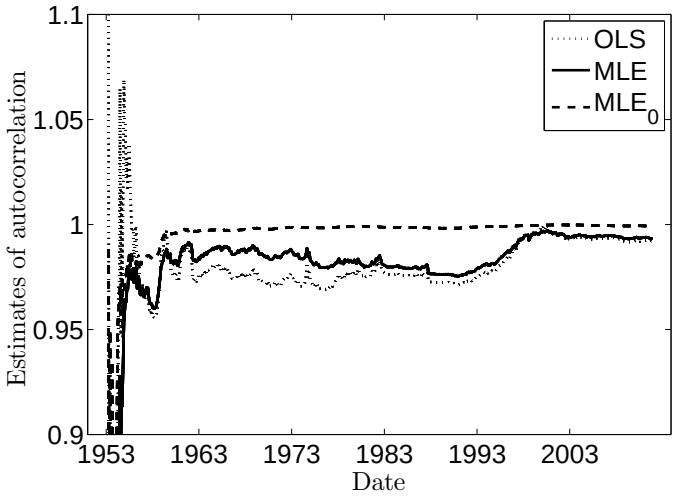


Fig. D.5. For each month, beginning in January 1953, we estimate the autocorrelation coefficient of the dividend-price ratio using maximum likelihood (MLE), maximum likelihood with the restriction $\beta = 0$ (MLE_0), and Ordinary Least Squares (OLS), using data from January 1953 up until that month.