NBER WORKING PAPER SERIES

MAXIMUM LIKELIHOOD ESTIMATION OF THE EQUITY PREMIUM

Efstathios Avdis
Jessica A. Wachter

Working Paper 19684
http://www.nber.org/papers/w19684

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2013

We are grateful to Kenneth Ahern, John Campbell, John Cochrane, Frank Diebold, Greg Duffee, Ian Dew- Becker, Adlai Fisher, Robert Hall, Soohun Kim, Ilaria Piatti, Jonathan Wright, Motohiro Yogo and seminar participants at the University of Alberta, the Wharton School, the NBER Forecasting & Empirical Methods Workshop, the SFS Cavalcade, the SoFiE Conference and the EFA Conference for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Maximum likelihood estimation of the equity premium
Efstathios Avdis and Jessica A. Wachter
NBER Working Paper No. 19684
November 2013, Revised September 2015
JEL No. C32,C58,G11,G12

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 tighter estimates under a range of specifications.

Efstathios Avdis
University of Alberta
3-30B Business
Edmonton, AB
Canada T6G 2R6
avdis@ualberta.ca

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

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 usually 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. If one is computing the sample average, a tighter estimate can be obtained only by extending the data series in time which has the disadvantage that the data are potentially less relevant to the present day.

Given the challenge in estimating sample means, it is not surprising that a number of studies investigate how to estimate the equity premium using techniques other than taking the sample average. These include making use of survey evidence (Claus and Thomas, 2001; Graham and Harvey, 2005; Welch, 2000), data on the cross section (Polk, Thompson, and Vuolteenaho, 2006), and data on stock return volatility (Pástor and Stambaugh, 2001). The branch of the literature most closely related to our work uses the accounting identity that links prices, dividends, and returns (Blanchard, 1993; Constantinides, 2002; Fama and French, 2002; Donaldson, Kamstra, and Kramer, 2010). The idea is simple in principle, but the implementation is inherently complicated by

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

the fact that the formula for returns is additive, while incorporating estimates of future dividend growth requires multi-year discount rates which are multiplicative.[2] As DeLong and Magin (2009) discuss in a survey of the literature, it is not clear why such methods would necessarily improve the estimation of the equity premium.

In this paper, we propose a method of estimating the equity premium that incorporates additional information contained in the time series of prices and dividends in a simple and econometrically-motivated way. Like the papers above, our work relies on a 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%. In annual terms, these translate to 5.2% and 3.9% respectively. Assuming that returns are approximately log-normally distributed, we can also derive implications for the equity premium computed in levels: in monthly terms the sample average implies an equity premium of 0.53%, or 6.37% per annum, while maximum likelihood implies an equity premium of 0.42% per month, or 5.06% per annum.

Besides showing that our method yields economically significant differences, we also perform a Monte Carlo experiment to demonstrate that, in finite samples and under a number of different assumptions on the data generating process, the maximum likelihood method is substantially less noisy than the sample average. For example, under our benchmark simulation, the sam-

[2]Fama and French (2002) have a relatively simple implementation in which they replace price appreciation by dividend growth in the expected return equation. We will discuss their paper in more detail in what follows.

ple average has a standard error of 0.089%, while our estimator has a standard error of only 0.050%.

Further, we derive formulas that give the 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. Thus maximum likelihood implies an equity premium that is below the sample average. Not surprisingly, given this intuition, we show by Monte Carlo simulations that the effect of our procedure is stronger, the more persistent the predictor variable.

The remainder of our paper proceeds as follows. Section 2 describes our statistical model and estimation procedure. Section 3 describes our results. Section 4 describes the intuition for our efficiency results and how these results depend on the parameters of the data generating process. Section 5 shows the applicability of our procedure under alternative data generating processes. First, we show how to adapt our procedure to account for conditional heteroskedasticity. Second, we consider the performance of our estimation procedure from Section 2 when the likelihood function is mis-specified in important ways. Third, we consider the implications of structural breaks for our analysis. Section 6 concludes.

2 Statistical Model and Estimation

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 $\theta < 1$; later we discuss the implications of relaxing this assumption. Note that our assumptions on the shocks imply that μ_r is the equity premium and that μ_x is the mean of x_t. While we focus on the case that the shocks are normally distributed and iid, we also explore robustness to alternative distributional assumptions.

Equations (1a) and (1b) for the return and predictor processes are standard in the literature. Indeed, the equation for returns 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.[3] Indeed,

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

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.[4] Our work differs from this literature in that μ_r is not a nuisance parameter but rather the focus of our study.

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

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

Maximizing this likelihood function is equivalent to running ordinary least squares regression. 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),

[4]See for example Bekaert, Hodrick, and Marshall (1997), Campbell and Yogo (2006), Nelson and Kim (1993), and Stambaugh (1999) for discussions on the bias in estimation of β and Cavanagh, Elliott, and Stock (1995), Elliott and Stock (1994), Jansson and Moreira (2006), Torous, Valkanov, and Yan (2004) and Ferson, Sarkissian, and Simin (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).

assuming that it is a draw from the stationary distribution. In our case, the 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

$$p\left(r_1,\ldots,r_T;x_0,\ldots,x_T|\mu_r,\mu_x,\beta,\theta,\Sigma\right) = \left(2\pi\sigma_x^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{x_0-\mu_x}{\sigma_x}\right)^2\right\}\times$$
$$|2\pi\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\sigma_v^2}{|\Sigma|}\sum_{t=1}^{T}u_t^2 - 2\frac{\sigma_{uv}}{|\Sigma|}\sum_{t=1}^{T}u_t v_t + \frac{\sigma_u^2}{|\Sigma|}\sum_{t=1}^{T}v_t^2\right)\right\}. \quad (3)$$

We follow Box and Tiao in referring to (2) as the conditional likelihood and (3) as the exact likelihood. Recent work that makes use of the exact likelihood in predictive regressions includes Stambaugh (1999) and Wachter and Warusawitharana (2009, 2012), who focus on estimation of the predictive coefficient β.[5] Other previous studies have focused on the effect of incorporating this first term (referred to as the initial condition) on unit root tests (Elliott, 1999; Müller and Elliott, 2003).[6]

We derive the values of μ_r, μ_x, β, θ, σ_u^2, σ_v^2 and σ_{uv} that maximize the likelihood (3) 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. In what follows, we refer to this procedure as maximum likelihood

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

[6]We could extend our results to multiple predictor variables (Kelly and Pruitt (2013), for example, allow multiple valuation ratios to predict returns), though to keep this manuscript of manageable size, we do not do so here. The likelihood function in (3) admits a generalization to multiple predictors, as can be found in Hamilton (1994).

estimation (MLE) even when we examine cases in which it is mis-specified. Depending on the context, we may also refer to it as our benchmark procedure.

In this paper, we compare estimating the equity premium using maximum likelihood versus the sample mean.[7] 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 (3) 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.

This reasoning holds asymptotically as the sample size grows large. Several practical considerations might be expected to work against this reasoning in finite samples. First, one might ask whether maximum likelihood delivers a substantively different, and more reliable, estimator than the sample mean. The asymptotic results say 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

[7]Another point of comparison is the estimate of the mean return from the conditional likelihood (2). Simulation results show that this estimator is less efficient than both the estimator from the exact likelihood and the sample mean. Additional information in regards to this estimator is available from the authors upon request.

could easily be outweighed in practice by the need to estimate a more complicated system. Finally, estimation of the 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

We calculate maximum likelihood estimates of the parameters in our predictive system for the excess return of the value-weighted market portfolio from CRSP. Recall that our object of interest is r_t, the logarithm of the gross return in excess of the riskfree asset: $r_t = \log(1 + R_t) - \log(1 + R_t^f)$. We take R_t to be the monthly net return of the value-weighted market portfolio and R_t^f to be the monthly net return of the 30-day Treasury Bill. We use 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.

3 Results

3.1 Point estimates

Table 1 reports estimates of the parameters of our statistical model given in (1). We report estimates for the 1927-2011 sample and for the 1953-2011 postwar subsample. For the postwar subsample, the equity premium from MLE is 0.322% in monthly terms and 3.86% per annum. In contrast, the sample average (given under the column labeled "OLS") is 0.433% in monthly terms, or 5.20% 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.

Table 1 also reports results for maximum likelihood estimation of the predictive coefficient β, the autoregressive coefficient θ, and the standard deviations and correlation between the shocks. The estimation of the standard deviations and correlation are nearly identical across the two methods, not surprisingly, because these can be estimated precisely in monthly data. Estimates for the average value of the predictor variable, the predictive coefficient and the autoregressive coefficient are noticeably different. The estimate for the average of the predictor variable is lower for maximum likelihood estimation (MLE) than for OLS in both samples. The difference in the postwar data is 4 basis points, 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.

3.2 Efficiency

We now return to the question of efficiency. We ask, does our maximum likelihood procedure reduce estimation noise in finite samples? We simulate 10,000 samples of excess returns and predictor variables, each of length equal to the data. Namely, we simulate from (1), setting parameter values equal to their maximum likelihood estimates, and, for each sample, initializing x using a draw from the stationary distribution. For each simulated sample, we calculate sample averages, OLS estimates and maximum likelihood estimates, generating a distribution of these estimates over the 10,000 paths.[8]

Table 2 (Panel A) reports the means, standard deviations, and the 5th, 50th, and 95th percentile values of a simulation calibrated using the postwar sample. While the sample average of the excess return has a standard devi-

[8]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. See Appendix A for further discussion of the polynomial for θ.

ation of 0.089, the maximum likelihood estimate has a standard deviation of only 0.050 (unless stated otherwise, units are in monthly percentage terms). [9] Besides lower standard deviations, 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.

Table 2 also shows that the maximum likelihood estimate of the mean of the predictor has a lower standard deviation 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 deviation and confidence intervals 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, in terms of the parameters of this system at least, the equity premium is unique in the improvement offered by maximum likelihood. This is in part due to the fact that estimation of first moments is more difficult than that of second moments in the time series (Merton, 1980). However, the result that the mean of returns is affected more than the mean of the predictor shows that this is not all that is going on. We return to this issue in Section 4.

Figure 1 provides another view of the difference between the sample mean and the maximum likelihood estimate of the equity premium. The solid line shows the probability density of the maximum likelihood estimates while the dashed line shows the probability density of the sample mean.[10] The data gen-

[9] Table A.1 shows an economically significant decline in standard deviation for the long sample as well: the standard deviation falls from 0.080 to 0.058. It is noteworthy that our results still hold in the longer sample, indicating that our method has value even when there is a large amount of data available to estimate the sample mean.

[10] Both densities are computed non-parametrically and smoothed by a normal kernel.

erating process is calibrated to the postwar period, assuming the parameters estimated using maximum likelihood (unless otherwise stated, all simulations that follow assume this calibration). The distribution of the maximum likelihood estimate is visibly more concentrated around the true value of the equity premium, and the tails of this distribution fall well under the tails of the distribution of sample means.[11] For the remainder of the paper, we refer to this data generating process, namely (1) with parameters given by maximum likelihood estimates from the postwar sample, as our benchmark case. Unless otherwise specified, we simulate samples of length equal to the postwar sample in the data (707 months).

It is well known that OLS estimates of predictive coefficients can be biased (Stambaugh, 1999). Panel A of Table 2 replicates this result: the "true" value of the predictive coefficient β in the simulated data is 0.69, however, 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

[11] In Table 2, we used coefficients estimated by maximum likelihood to evaluate whether MLE is more efficient than OLS. Perhaps it is not surprising that MLE delivers better estimates, if we use the maximum likelihood estimates themselves in the simulation. However, Table A.3 shows nearly identical results from setting the parameters equal to their sample means and OLS estimates. We perform more extensive robustness checks in Section 5.

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 if 0.072. Namely, 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.[12] We will refer to this as our benchmark case with bias-correction.

3.3 The equity premium in levels

So far we have defined the equity premium in terms of log returns. However, our result is also indicative of a lower equity premium using return levels. For simplicity, assume that the log returns $\log(1+R_t)$ are normally distributed. Then

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

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

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

[12] Table A.2 shows results under bias correction and fat-tailed shocks. Our results are virtually unchanged.

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.[13] Thus our estimate of the risk premium in return levels is 131 basis lower than taking the sample average, in line with our results for log returns.

3.4 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:

$$R_t = \frac{D_t}{P_{t-1}} + \frac{P_t - P_{t-1}}{P_{t-1}},$$

and taking the expectation:

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

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

$$d_{t+1} - d_t = x_{t+1} - x_t + p_{t+1} - p_t. \tag{4}$$

[13] 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%.

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

$$E[d_{t+1} - d_t] = E[p_{t+1} - p_t]. \tag{5}$$

However, exponentiating (4) and subtracting 1 implies

$$\frac{D_{t+1} - D_t}{D_t} = \frac{X_{t+1}}{X_t}\frac{P_{t+1}}{P_t} - 1. \tag{6}$$

That is, stationarity of X_t implies (5), 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, (6) 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 $E[X_{t+1}/X_t] > 1$.[14] 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.[15] Indeed, the intuition that we will develop in the next section is closely related to

[14]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:

$$\mathrm{Var}(d_{t+1} - d_t) = \mathrm{Var}(p_{t+1} - p_t). \tag{7}$$

To see this, note that (5), combined with log-normality, implies that

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

If (7) 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).

[15]This point is also made by Constantinides (2002), who suggests adjusting the mean

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.

4 Discussion

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

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 mean return.

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 (2). 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.[16]

[16]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 entirely representative of values of the predictor variable in population. Namely, the values of the predictor variable for the last T observations 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.[17] Note that this result operates purely through the correlation of the shocks, and is not related to predictability.[18]

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 A.2 shows that indeed these terms tend to have opposite signs in the simulated data.[19]

[17] 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.

[18] Ultimately, however, there may be a connection in that variation in the equity premium is the main driver of variation in the dividend-price ratio and thus the reason why the shocks are negatively correlated.

[19] There is a small opposing effect on the sign of the predictability term. Note that the

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 2 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.

4.2 Properties of the maximum likelihood estimator

In this section we investigate the properties of the maximum likelihood estimator, and, in particular, how the variance of the estimator depends on the persistence of the predictor variable, the amount of predictability, and the correlation between the shocks to the predictor and the shocks to returns.

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

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.

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 2, 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 B).

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). 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 Variance of estimator under alternative parameter assumptions

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. In the historical sample, predictability works against us in finding a lower equity premium. Indeed, 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 2 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 2). 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.[20] 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

[20]Wachter and Warusawitharana (2015) show in a Bayesian setting that, if one holds a belief that there is no predictability, the posterior distribution for the autoregressive coefficient shifts upward towards unity. Cochrane (2008) makes an analogous point using frequentist methods.

[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.

ratio is reliably estimated to be strongly negative.[23] Nonetheless, it is a stark illustration of the conditions under which our efficiency gains break down.

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

As is well-known, stock returns exhibit time-varying volatility (French, Schwert, and Stambaugh, 1987; Bollerslev, Chou, and Kroner, 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 sam-

[23]It does suggest, however, that including data on predictor variables that have low persistence and/or low realized correlations with returns will not impact estimates of the equity premium nearly to the extent of the dividend-price ratio.

ples? (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 there-

[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.

fore

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

This likelihood function conditions on x_0, and thus is the GARCH analogue of the conditional maximum likelihood function (2). 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

[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 A.4. Similarly to previous studies (e.g. French, Schwert, and Stambaugh (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 3 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 2). 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.

5.2.1 The random walk model

Given the observed high autocorrelation of the dividend-price ratio, a natural extension is to consider a random walk. 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.[27] 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

[27] 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.

likelihood procedure is mis-specified because it assumes stationarity and allows for predictability. Of course assumptions of OLS are also violated, as discussed above.

Table 4 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. As previously discussed, the predictive coefficient and the autoregressive coefficient are biased upward and downward respectively, and this is clearly shown in the table. As a result, maximum likelihood still identifies a positive β and a stationary dividend-price ratio, even though these are not the characteristics of the data generating process.

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.

5.2.2 Predictor with Time Trend

The previous section shows that our method can still be effective under a random-walk model for the dividend-price ratio. What about other forms of non-stationarity? Using the intuition from Section 4, we can reason backwards to find a model seems particularly likely to cause problems for our estimation method. Such a model would lead our method to conclude that the average shock is non-zero more often than it is.

These considerations lead us to consider 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. We consider a calibration of (18) that both fits the data, and represents a worst-case scenario from the point of view of our method. 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 5. 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: 0280% versus 0.322%. The maximum likelihood estimator has a lower standard error, but this doesn't matter because it is in fact an inconsistent estimator for the equity premium.

Why does the maximum likelihood estimator fail in this case? Consider first the estimation of the process for x_t. The true mean of x_t is undefined. However, in every sample there will be an observed mean. This sample mean will be on average lower than the true value of μ_x because the time trend lowers the level of the dividend-price ratio. The MLE will be slightly higher than the sample mean because it will correct for what it sees as an unusual

series of shocks. 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.[28]

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

[28]Hansen, Heaton, and Li (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.

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 6. 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 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 7 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 3.2, 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 indicate that the small-sample noise is greatly reduced.

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, Takeshi, 1985, *Advanced Econometrics*. (Harvard University Press Cambridge, MA).

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

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

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

Bekaert, Geert, Robert J. Hodrick, and David A. Marshall, 1997, On biases in tests of the expectations hypothesis of the term structure of interest rates, *Journal of Financial Economics* 44, 309–348.

Blanchard, Olivier J., 1993, Movements in the Equity Premium, *Brookings Papers on Economic Activity* 1993, 75–138.

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

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

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

Campbell, John Y., 2003, Consumption-based asset pricing, in G. Constantinides, M. Harris, and R. Stulz, eds.: *Handbook of the Economics of Finance, vol. 1b* (Elsevier Science, North-Holland).

Campbell, John Y., 2006, Household Finance, *Journal of Finance* 61, 1553 – 1604.

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

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

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

Cavanagh, Christopher L., Graham Elliott, and James H. Stock, 1995, Inference in models with nearly integrated regressors, *Econometric Theory* 11, 1131–1147.

Claus, James, and Jacob Thomas, 2001, Equity Premia as Low as Three Percent? Evidence from Analysts' Earnings Forecasts for Domestic and International Stock Markets, *The Journal of Finance* 56, 1629–1666.

Cochrane, John H., 2008, The Dog That Did Not Bark: A Defense of Return Predictability, *The Review of Financial Studies* 21, 1533–1575.

Constantinides, George M., 2002, Rational Asset Prices, *The Journal of Finance* 57, 1567–1591.

DeLong, J. Bradford, and Konstantin Magin, 2009, The U.S. Equity Return Premium: Past, Present, and Future, *The Journal of Economic Perspectives* 23, 193–208.

Diebold, Francis X., and Til Schuermann, 2000, Exact maximum likelihood estimation of observation-driven econometric models, in M. Weeks R.S. Mariano, and T. Schuermann, eds.: *Simulation-Based Inference in Econometrics: Methods and Applications* (Cambridge University Press,).

Donaldson, R. Glen, Mark J. Kamstra, and Lisa A. Kramer, 2010, Estimating the equity premium, *Journal of Financial and Quantitative Analysis* 45, 813–846.

Elliott, Graham, 1999, Efficient Tests for a Unit Root When the Initial Observation is Drawn from Its Unconditional Distribution, *International Economic Review* 40, 767–783.

Elliott, Graham, and James H Stock, 1994, Inference in Time Series Regression When the Order of Integration of a Regressor Is Unknown, *Econometric Theory* 10, 672–700.

Fama, Eugene F., and Kenneth R. French, 1989, Business conditions and expected returns on stocks and bonds, *Journal of Financial Economics* 25, 23–49.

Fama, Eugene F., and Kenneth R. French, 2002, The Equity Premium, *The Journal of Finance* 57, pp. 637–659.

Ferson, Wayne E., Sergei Sarkissian, and Timothy T. Simin, 2003, Spurious regressions in financial economics?, *Journal of Finance* 58, 1393–1413.

French, Kenneth R., G. William Schwert, and Robert F. Stambaugh, 1987, Expected stock returns and volatility, *Journal of Financial Economics* 19, 3–29.

Graham, John R., and Campbell R. Harvey, 2005, The long-run equity risk premium, *Finance Research Letters* 2, 185–194.

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

Hansen, Lars Peter, John C. Heaton, and Nan Li, 2008, Consumption strikes back? Measuring long run risk, *Journal of Political Economy* 116, 260–302.

Jansson, Michael, and Marcelo J. Moreira, 2006, Optimal Inference in Regression Models with Nearly Integrated Regressors, *Econometrica* 74, 681–714.

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

Kelly, Bryan, and Seth Pruitt, 2013, Market expectations in the cross-section of present values, *The Journal of Finance* 68, 1721–1756.

Kocherlakota, Narayana R., 1996, The Equity Premium: It's Still a Puzzle, *Journal of Economic Literature* 34, 42–71.

Lettau, Martin, and Stijn Van Nieuwerburgh, 2008, Reconciling the return predictability evidence, *Review of Financial Studies* 21, 1607–1652.

Lettau, Martin, and Jessica A. Wachter, 2007, Why is long-horizon equity less risky? A duration-based explanation of the value premium, *Journal of Finance* 62, 55–92.

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

Lynch, Anthony W., and Jessica A. Wachter, 2013, Using Samples of Unequal Length in Generalized Method of Moments Estimation, *Journal of Financial and Quantitative Analysis* 48, 277–307.

Malkiel, Burton Gordon, 2003, *A random walk down Wall Street*. (W. W. Norton and Company, Inc. New York, NY).

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

Mehra, Rajnish, and Edward C. Prescott, 2003, The equity premium in retrospect, in G. M. Constantinides, M. Harris, and R. M. Stulz, eds.: *Handbook of the Economics of Finance* (Elsevier, North-Holland).

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

Müller, Ulrich K., and Graham Elliott, 2003, Tests for Unit Roots and the Initial Condition, *Econometrica* 71, 1269–1286.

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

Pástor, Ľuboš, and Robert F. Stambaugh, 2001, The Equity Premium and Structural Breaks, *The Journal of Finance* 56, 1207–1239.

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

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

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

Polk, Christopher, Samuel Thompson, and Tuomo Vuolteenaho, 2006, Cross-sectional forecasts of the equity premium, *Journal of Financial Economics* 81, 101–141.

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

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

Singleton, Kenneth, 2006, *Empirical dynamic asset pricing: Model specification and econometric assessment*. (Princeton University Press Princeton, NJ).

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

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

Torous, Walter, Rossen Valkanov, and Shu Yan, 2004, On predicting stock returns with nearly integrated explanatory variables, *Journal of Business* 77, 937–966.

Van Binsbergen, Jules H., and Ralph S. J. Koijen, 2010, Predictive regressions: A present-value approach, *The Journal of Finance* 65, 1439–1471.

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

Wachter, Jessica A., and Missaka Warusawitharana, 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, Ivo, 2000, Views of Financial Economists on the Equity Premium and on Professional Controversies, *The Journal of Business* 73, 501–537.

Table 1: Maximum Likelihood and OLS Estimates

	1953–2011		1927–2011	
	OLS	MLE	OLS	MLE
μ_r	0.433	0.322	0.464	0.391
μ_x	−3.545	−3.504	−3.374	−3.383
β	0.828	0.686	0.623	0.650
θ	0.992	0.993	0.992	0.991
σ_u	4.414	4.416	5.466	5.464
σ_v	0.046	0.046	0.057	0.057
ρ_{uv}	−0.961	−0.961	−0.953	−0.953

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 standard deviations σ_u and σ_v and 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. Data are monthly. Means and standard deviations of returns are in percentage terms. Under 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. Under the MLE columns, parameters are estimated using maximum likelihood.

Table 2: Small-sample distribution of estimated parameters

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
Panel A: DGP calibrated to maximum likelihood estimates							
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: DGP calibrated to bias-corrected estimates							
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.

Table 3: 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 4: 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	undefined	Sample	−2.988	0.699	−4.130	−2.996	−1.845
		MLE	−2.986	0.637	−4.006	−2.997	−1.971
β	0	OLS	0.710	0.608	0.005	0.571	1.883
		MLE	0.629	0.562	0.062	0.467	1.729
θ	1.000	OLS	0.992	0.006	0.980	0.994	1.000
		MLE	0.993	0.006	0.981	0.995	0.999
σ_u	4.423	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}	−0.962	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 5: 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.

Table 6: Sub-sample estimates

	1953–1994		1995–2011	
	OLS	MLE	OLS	MLE
μ_r	0.439	0.315	0.411	0.336
μ_x	−3.342	−3.337	−4.048	−3.955
β	2.538	2.186	2.614	1.968
θ	0.977	0.981	0.972	0.979
σ_u	4.205	4.210	4.840	4.842
σ_v	0.043	0.043	0.051	0.051
ρ_{uv}	−0.967	−0.967	−0.948	−0.949

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. Under 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. Under the MLE columns, parameters are estimated using maximum likelihood.

Table 7: Small-sample distribution of estimators in simulations calibrated to subsamples from Table 6

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95%
		Panel A1: DGP calibrated to 1953–1994 period					
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: DGP calibrated to 1953–1994 period with 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: DGP calibrated to 1995–2011 period					
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: DGP calibrated to 1995–2011 period with 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.

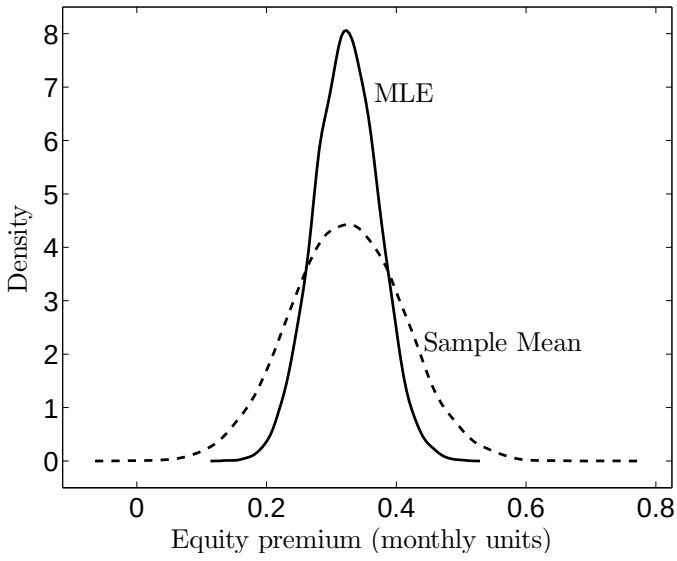

Figure 1: Densities of the estimators of the equity premium in repeated samples of length equal to the postwar data. The solid line shows the density of the maximum likelihood estimate while the dashed line shows the density of the sample mean.

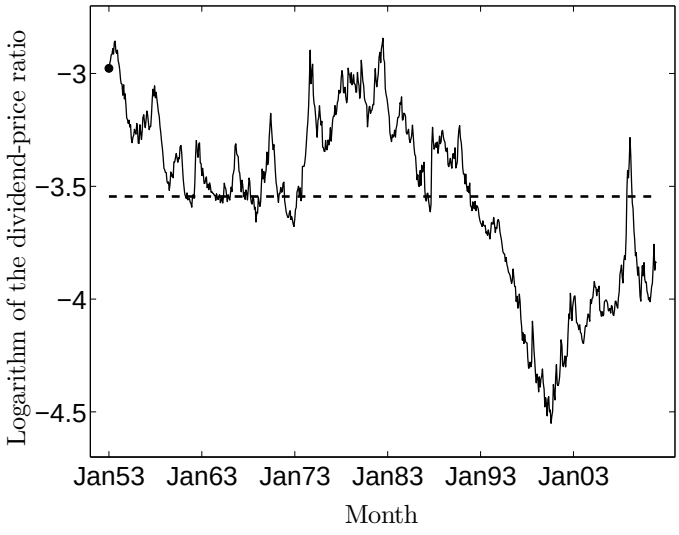

Figure 2: The logarithm of the dividend-price ratio over the period January 1953 to December 2011 (the postwar sample). The dotted line indicates the mean, and the black dot the initial value.

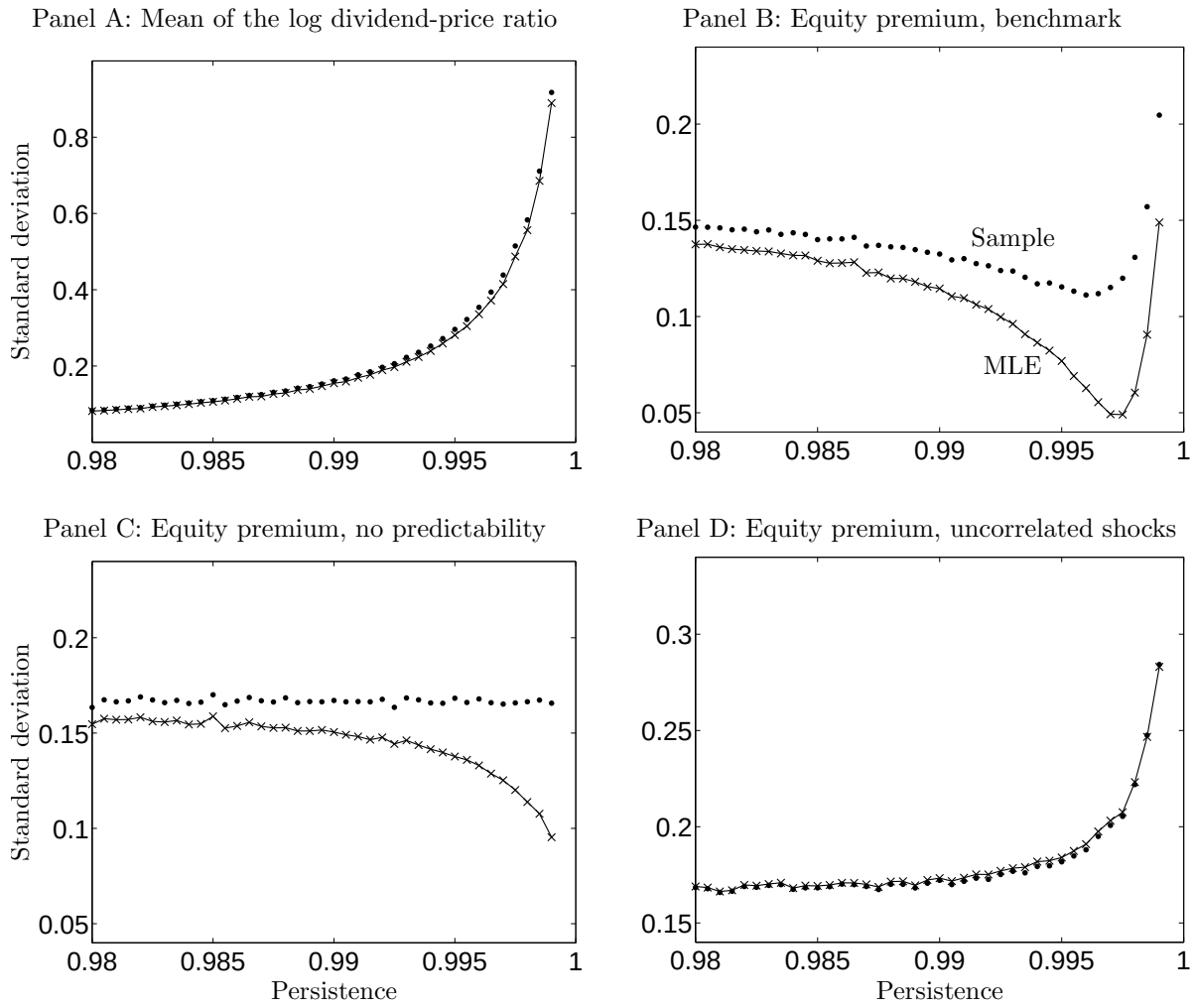

Figure 3: Standard deviation of estimators of the mean of the log-dividend price ratio (Panel A) and of the equity premium (Panels B–D). Estimators are the sample mean (dots) and maximum likelihood (crosses). For each value of the autocorrelation θ, we simulate 10,000 monthly samples and calculate the 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.

Appendix

A Derivation of the Maximum Likelihood Estimators

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 $\mathcal{L}$:

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

where $|\Sigma| = \sigma_u^2\sigma_v^2 - \sigma_{uv}^2$. The first-order conditions arise from setting the following partial derivatives of the likelihood function to zero:

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

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

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

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

$$0 = \frac{\partial}{\partial\sigma_{uv}}\mathcal{L} = -T\frac{2\sigma_{uv}}{|\Sigma|}$$

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

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

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

Define the residuals

$$\hat{u}_t = r_t - \hat{\mu}_r - \hat{\beta}(x_{t-1} - \hat{\mu}_x)$$
$$\hat{v}_t = x_t - \hat{\mu}_x - \hat{\theta}(x_{t-1} - \hat{\mu}_x).$$

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.1c) and (A.1d) gives

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

which we can write as

$$\hat{\mu}_x = \frac{\left(1+\hat{\theta}\right)x_0 + \sum_{t=1}^{T}\left(x_t - \hat{\theta}x_{t-1}\right)}{\left(1+\hat{\theta}\right) + \left(1-\hat{\theta}\right)T}. \tag{A.3}$$

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.1e), (A.1f) and (A.1g) give the relations

$$T\hat{\sigma}_u^2 = -\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2}\hat{\sigma}_{uv} + (1-\hat{\theta}^2)(x_0 - \hat{\mu}_x)^2\left(\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2}\right)^2 + \sum_{t=1}^{T}\hat{u}_t^2, \tag{A.4}$$

$$(T+1)\hat{\sigma}_v^2 = (1-\hat{\theta}^2)(x_0 - \hat{\mu}_x)^2 + \sum_{t=1}^{T}\hat{v}_t^2, \tag{A.5}$$

$$\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2} = \frac{\sum_{t=1}^{T}\hat{u}_t\hat{v}_t}{\sum_{t=1}^{T}\hat{v}_t^2}. \tag{A.6}$$

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.1a) and (A.1b) gives

$$0 = \sum_{t=1}^{T}(\hat{\mu}_x - x_{t-1})\hat{v}_t + \hat{\sigma}_v^2\frac{\hat{\theta}}{1-\hat{\theta}^2} - \hat{\theta}(x_0 - \hat{\mu}_x)^2. \tag{A.7}$$

Here $\hat{\mu}_x$ and $\hat{v}_t$ are functions of only $\hat{\theta}$ and the data, so if we combine (A.7) and (A.5) we can get an equation for $\hat{\theta}$:

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

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

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

and rearrange to obtain

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

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

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.3) gives the estimator for μ_x and (A.5) 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.1c) gives

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

Combining this with the first-order condition (A.1a) yields

$$\hat{\beta} = \beta^{\text{OLS}} + \frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2} \left(\hat{\theta} - \theta^{\text{OLS}} \right), \tag{A.9}$$

where

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

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

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

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

Equations (A.6), (A.8) and (A.9) 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

$$\hat{\mu}_r = \frac{1}{J}\left[\frac{1}{T}\sum_{t=1}^{T} r_t - \left(\frac{1}{T}\sum_{t=1}^{T} x_t - \hat{\mu}_x\right)\frac{F - \beta^{\text{OLS}}H}{1 + (\hat{\theta} - \theta^{\text{OLS}})H} - \left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1} - \hat{\mu}_x\right)\frac{\beta^{\text{OLS}}(1 + \hat{\theta}H) - \theta^{\text{OLS}}F}{1 + (\hat{\theta} - \theta^{\text{OLS}})H}\right] \tag{A.10}$$

$$\hat{\beta} = \frac{\beta^{\text{OLS}} + (\hat{\theta} - \theta^{\text{OLS}})F}{1 + (\hat{\theta} - \theta^{\text{OLS}})H} - \frac{(\hat{\theta} - \theta^{\text{OLS}})G}{1 + (\hat{\theta} - \theta^{\text{OLS}})H}\hat{\mu}_r \tag{A.11}$$

$$\frac{\hat{\sigma}_{uv}}{\hat{\sigma}_v^2} = \frac{F - \beta^{\text{OLS}}H}{1 + (\hat{\theta} - \theta^{\text{OLS}})H} - \frac{G}{1 + (\hat{\theta} - \theta^{\text{OLS}})H}\hat{\mu}_r, \tag{A.12}$$

where

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

$$F = \frac{\sum_{t=1}^{T} r_t \hat{v}_t}{\sum_{t=1}^{T} \hat{v}_t^2}$$

$$G = \frac{\sum_{t=1}^{T} \hat{v}_t}{\sum_{t=1}^{T} \hat{v}_t^2}$$

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

Expressions (A.10) and (A.11) 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.12) gives the estimator the estimator of σ_{uv} via (A.5), which further yields the estimator of σ_u^2 via (A.4).

B 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):

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

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

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

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 (B.1), $\theta^{k-1}\beta\sigma_{uv}$ 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

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

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,

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

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, $\text{Cov}\,(r_t, r_{t+k}) < 0$ whenever $\theta < \theta_2$ and $\text{Cov}\,(r_t, r_{t+k}) > 0$ whenever $\theta > \theta_2$.

C The Variance of the Sample Mean Return

By definition

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

thus

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

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,

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} u_t\right) &= \sigma_u^2 \frac{1}{T}, \\
\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}\right) &= \frac{\sigma_v^2}{1-\theta^2}\left[\frac{1}{T}\left(1 + 2\frac{\theta}{1-\theta}\right) + \frac{2}{T^2}\frac{\theta(\theta^T - 1)}{(1-\theta)^2}\right], \\
\mathrm{Cov}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t-1}, \frac{1}{T}\sum_{t=1}^{T} u_t\right) &= \sigma_{uv}\left[\frac{1}{T}\frac{1}{1-\theta} + \frac{1}{T^2}\frac{\theta^T - 1}{(1-\theta)^2}\right],
\end{aligned}$$

so that

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

It follows that

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

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.

Table A.1: 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 A.2: 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 A.3: Small-sample distribution of estimators: Calibration to OLS estimates

	True Value	Method	Mean	Std. Dev.	5 %	50 %	95 %
		Panel A: January 1953 to December 2011					
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
		Panel B: January 1927 to December 2011					
μ_r	0.464	Sample	0.463	0.082	0.326	0.462	0.596
		MLE	0.464	0.058	0.367	0.463	0.560
μ_x	−3.374	Sample	−3.373	0.200	−3.702	−3.373	−3.044
		MLE	−3.373	0.194	−3.690	−3.374	−3.054
β	0.623	OLS	1.019	0.543	0.322	0.925	2.051
		MLE	0.995	0.527	0.329	0.903	1.983
θ	0.992	OLS	0.987	0.006	0.976	0.988	0.995
		MLE	0.988	0.006	0.977	0.989	0.995
σ_u	5.466	OLS	5.465	0.121	5.269	5.463	5.668
		MLE	5.463	0.121	5.268	5.461	5.666
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 as in postwar data. Parameters are set to their OLS estimates. 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 A.4: 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.

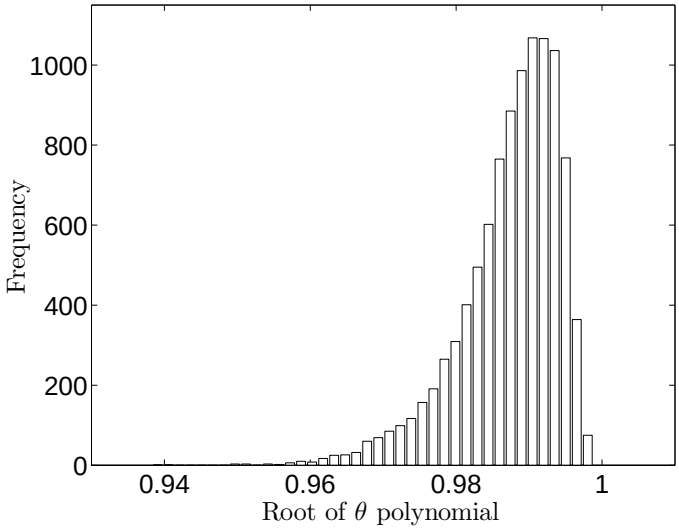

Figure A.1: Histogram of maximum likelihood estimates of θ, the autocorrelation of the dividend-price ratio from simulated data. We simulate 10,000 monthly data samples from (1) with length and parameters as in the postwar data series.

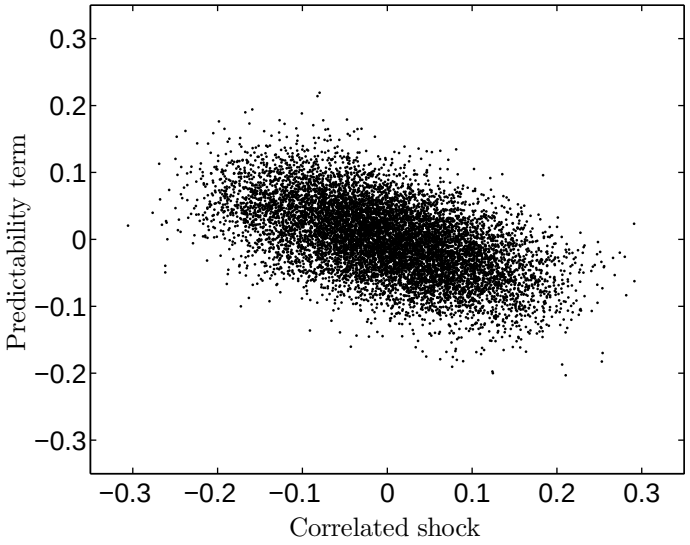

Figure A.2: We simulate 10,000 monthly data samples from (1) with length and parameters as in the postwar data series. The figure shows the joint distribution of the predictability term $\hat{\beta}\frac{1}{T}\sum_{t=1}^{T}(x_{t-1}-\hat{\mu}_x)$ and the correlated shock term $\frac{1}{T}\sum_{t=1}^{T}\hat{u}_t$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