What is the Expected Return on the Market?

Ian Martin*

June, 2015

Abstract

This paper presents a new lower bound on the equity premium in terms of a volatility index, SVIX, that can be calculated from index option prices. This bound, which relies only on very weak assumptions, implies that the equity premium is extremely volatile, and that it rose above 20% at the height of the crisis in 2008. More aggressively, I argue that the lower bound—whose time-series average is about 5%—is approximately tight and that the high equity premia available at times of stress largely reflect high expected returns over the very short run. Under a stronger assumption, I show how to use option prices to measure the probability that the market goes up (or down) over some given horizon, and to compute the expected excess return on the market conditional on the market going up (or down).

*London School of Economics; http://personal.lse.ac.uk/martiniw. I am grateful to John Campbell, John Cochrane, George Constantinides, Darrell Duffie, Bernard Dumas, Lars Hansen, Stefan Nagel, Christopher Polk, José Scheinkman, Dimitri Vayanos, and to seminar participants at Stanford University, Northwestern University, the NBER Summer Institute, INSEAD, Swiss Finance Institute, MIT Sloan, Harvard University, Morgan Stanley, Princeton University, London School of Economics, the Federal Reserve Bank of Atlanta, Toulouse School of Economics, FIME, BI Business School, Copenhagen Business School, Washington University in St Louis, Stockholm School of Economics, the Brazilian Finance Society, Tuck School of Business, the University of Chicago and Warwick Business School for their comments.

The expected excess return on the market, or equity premium, is one of the central quantities of finance. Aside from its obvious intrinsic interest, the equity premium is a key determinant of the risk premium required for arbitrary assets in the CAPM and its descendants; and time-variation in the equity premium lies at the heart of the literature on excess volatility.

The starting point of this paper is an identity that relates the market's expected return to its risk-neutral variance. Under the weak assumption of no-arbitrage, the latter can be measured unambiguously from index option prices. I call the associated volatility index SVIX and use the identity (coupled with a minimal assumption, the *negative correlation condition*, introduced in Section 1) to derive a lower bound on the equity premium in terms of the SVIX index. The bound implies that the equity premium is extremely volatile, and that it rose above 21% at the height of the crisis in 2008. At horizons of less than a year, the equity premium fluctuates even more wildly: the lower bound on the *monthly* equity premium exceeded 4.5% (unannualized) in November 2008.

I go on to argue, more aggressively, that the lower bound appears empirically to be approximately tight, so that the SVIX index provides a direct measure of the equity premium. While it is now well understood that the equity premium is time-varying, this paper deviates from the literature in its basic aim, which is to use theory to motivate a signal of whether expected returns are high or low at a given point in time that is based *directly on asset prices*. The distinctive features of my approach, relative to the literature, are that (i) the predictor variable, SVIX^2, is motivated by asset pricing theory; (ii) no parameter estimation is required, so concerns over in-sample/out-of-sample fit do not arise; and (iii) since the SVIX^2 index is an asset price, I avoid the need to use infrequently-updated accounting data. My approach therefore allows the equity premium to be measured *in real time*.

The SVIX^2 index can be interpreted as the equity premium perceived by an unconstrained rational investor with log utility who is fully invested in the market. This is a sensible benchmark even if there are many investors who are constrained and many investors who are irrational, and it makes for a natural comparison with survey evidence on investor expectations, as studied by Shiller (1987), Ben-David, Graham and Harvey (2013), and Greenwood and Shleifer (2014), among others. In particular, Greenwood and Shleifer (2014) emphasize that the 'return expectations' extracted

from surveys are negatively correlated with subsequent realized returns. In contrast, my measure of expected returns is positively correlated with subsequent returns—a minimal requirement for a measure of rationally expected returns.

I then compare the view of the equity premium that emerges from the SVIX measure to the conventional view based on valuation-ratio-based measures. (I take Campbell and Thompson (2008) as representative of the vast predictability literature[1] because their approach, like mine, avoids the in-sample/out-of-sample critique of Goyal and Welch (2008).) My results deviate from this conventional view in several interesting ways. I argue that the equity premium is more volatile, more right-skewed (particularly at short horizons), and that it fluctuates at a higher frequency than the literature has acknowledged.

I sharpen the distinction between the SVIX and valuation-ratio views of the world by focussing on two periods in which their predictions diverge. Valuation-ratio-based measures of the equity premium were famously bearish throughout the late 1990s (and as noted by Ang and Bekaert (2007) and Goyal and Welch (2008), that prediction is partially responsible for the poor performance of valuation-ratio predictors in recent years); in contrast, the SVIX index suggests that, at horizons up to one year, expected returns were high in the late 1990s. I argue that this distinction reflects the fact that valuation ratios should be thought of as predictors of very long run returns, whereas the SVIX index aims to measure short-run expected returns. I also show that (from the perspective of the unconstrained log investor whose expected equity premium is measured by SVIX^2) the expected return on the market *conditional on a market decline* was unusually low in the late 1990s, suggesting that sophisticated investors believed that a crash, if it occurred, might be severe; this is reminiscent of the findings of Brunnermeier and Nagel (2004). The most striking divergence in predictions, however, occurs on one of the most dramatic days in stock market history: the great crash of October 1987. On the valuation-ratio view of the world, the equity premium barely changed on Black Monday; on the SVIX view, it exploded.

[1]Early papers in this literature include Keim and Stambaugh (1986), Campbell and Shiller (1988), and Fama and French (1988).

1 Expected returns and risk-neutral variance

If we use asterisks to denote quantities calculated with risk-neutral probabilities, and M_T to denote the stochastic discount factor (SDF) that prices time-T payoffs from the perspective of time t, then we can price any time-T payoff X_T either via the SDF or by computing expectations with risk-neutral probabilities and discounting at the (gross) riskless rate, $R_{f,t}$, which is known at time t. The SDF notation,

$$\text{time-}t \text{ price of a claim to } X_T \text{ at time } T = \mathbb{E}_t(M_T X_T), \tag{1}$$

is commonly used in equilibrium models or, more generally, whenever there is an emphasis on the real-world distribution (whether from the subjective perspective of an agent within a model, or from the 'objective' perspective of the econometrician).

The risk-neutral notation,

$$\text{time-}t \text{ price of a claim to } X_T \text{ at time } T = \frac{1}{R_{f,t}} \mathbb{E}_t^* X_T, \tag{2}$$

is commonly used in derivative pricing, or more generally whenever the underlying logic is that of no-arbitrage. The choice of whether to use SDF or risk-neutral notation is largely a matter of taste; I will tend to follow convention by using the risk-neutral notation when no-arbitrage logic is emphasized.

Equations (1) and (2) can be used to translate between the two notations; thus, for example, the conditional risk-neutral variance of a gross return R_T is

$$\text{var}_t^* R_T = \mathbb{E}_t^* R_T^2 - (\mathbb{E}_t^* R_T)^2 = R_{f,t} \mathbb{E}_t \left(M_T R_T^2 \right) - R_{f,t}^2. \tag{3}$$

Expected returns and risk-neutral variance are linked by the following identity:

$$\begin{aligned} \mathbb{E}_t R_T - R_{f,t} &= \left[\mathbb{E}_t(M_T R_T^2) - R_{f,t} \right] - \left[\mathbb{E}_t(M_T R_T^2) - \mathbb{E}_t R_T \right] \\ &= \frac{1}{R_{f,t}} \text{var}_t^* R_T - \text{cov}_t(M_T R_T, R_T). \end{aligned} \tag{4}$$

The first equality adds and subtracts $\mathbb{E}_t(M_T R_T^2)$; the second exploits (3) and the fact that $\mathbb{E}_t M_T R_T = 1$.

The identity (4) decomposes the asset's risk premium into two components. It applies to any asset return R_T, but in this paper I will focus on the case in which R_T is the return on the S&P 500 index. In this case the first component, risk-neutral variance, can be computed directly given time-t prices of S&P 500 index options, as

will be shown in Section 3. The second component is a covariance term that can be controlled: under a weak condition (discussed in detail in Section 2), it is negative.

Definition 1. *Given a gross return R_T and stochastic discount factor M_T, the* negative correlation condition (NCC) *holds if* $\operatorname{cov}_t(M_T R_T, R_T) \leq 0$.

Together, the identity (4) and the NCC imply the following inequality, from which the results of the paper flow:

$$\mathbb{E}_t R_T - R_{f,t} \geq \frac{1}{R_{f,t}} \operatorname{var}_t^* R_T. \tag{5}$$

This is a bound in the opposite direction to the Hansen–Jagannathan (1991) bound. Together, the two bounds imply that

$$\frac{1}{R_{f,t}} \operatorname{var}_t^* R_T \leq \mathbb{E}_t R_T - R_{f,t} \leq R_{f,t} \cdot \sigma_t(M_T) \cdot \sigma_t(R_T),$$

where $\sigma_t(\cdot)$ denotes conditional (real-world) standard deviation. The left-hand inequality is (5). It has the advantage that it relates the unobservable equity premium to a *directly observable* quantity, risk-neutral variance; but the disadvantage that it requires the NCC to hold. In contrast, the right-hand inequality, the Hansen–Jagannathan bound, has the advantage of holding completely generally; but the disadvantage (noted by Hansen and Jagannathan) that it relates two quantities neither of which can be directly observed. Time-series averages must therefore be used as proxies for the true quantities of interest, forward-looking means and variances. This procedure requires assumptions about the stationarity and ergodicity of returns over appropriate sample periods and at the appropriate frequency. Such assumptions are not completely uncontroversial—see, for example, Malmendier and Nagel (2011).

Inequality (5) is reminiscent of the approach of Merton (1980), based on the equation

$$\text{instantaneous risk premium} = \gamma \sigma^2 \,, \tag{6}$$

where γ is a measure of aggregate risk aversion, and σ^2 is the instantaneous variance of the market return, and of a closely related calculation carried out by Cochrane (2011, p. 1082).

There are some important differences between the two approaches, however. The first is that Merton assumes that the level of the stock index follows a geometric Brownian motion, thereby ruling out the effects of skewness and of higher moments

by construction.[2] In contrast, we need no such assumption. Related to this, there is no distinction between risk-neutral and real-world (instantaneous) variance in a diffusion-based model: the two are identical, by Girsanov's theorem. Once we move beyond geometric Brownian motion, however, the appropriate generalization relates the risk premium to *risk-neutral* variance. As a bonus, this will have the considerable benefit that—unlike forward-looking real-world variance—forward-looking risk-neutral variance at time t can be directly and unambiguously computed from asset prices at time t, as I show in Section 3.

A second difference is that (6) requires that there is a representative agent with constant relative risk aversion γ. The NCC holds under considerably more general circumstances, as shown in Section 2.1.

Third, Merton implements (6) using realized historical volatility rather than by exploiting option price data, though he notes that volatility measures can be calculated "by 'inverting' the Black–Scholes option pricing formula." However, Black–Scholes implied volatility would only provide the correct measure of σ if we really lived in a Black–Scholes (1973) world in which prices followed geometric Brownian motions. The results of this paper show how to compute the right measure of variance in a more general environment.

2 The negative correlation condition

This section is devoted to arguing that the NCC holds when R_T is the return on the market. It is independent of the rest of the paper. I start by laying out some sufficient conditions for the NCC to hold in theoretical models. These sufficient conditions cover many of the leading macro-finance models, including Campbell and Cochrane (1999), Bansal and Yaron (2004), Bansal, Kiku, Shaliastovich and Yaron (2012), Campbell, Giglio, Polk and Turley (2012), Barro (2006), and Wachter (2013).[3]

I next test the plausibility of the NCC by carrying out a time-series estimation of two

[2]Cochrane's calculation also implicitly makes this assumption. Amongst others, Rubinstein (1973), Kraus and Litzenberger (1976), and Harvey and Siddique (2000) emphasize the importance of skewness in portfolio choice.

[3]In fact, I am not aware of any model that attempts to match the data quantitatively in which the NCC does not hold.

linear factor models: the three-factor model of Fama and French (1996), and the same model with a momentum factor included. Estimates of the covariance $\operatorname{cov}(M_T R_T, R_T)$ are negative, consistent with the NCC, and stable across sample periods. The estimates are close to zero both economically and statistically, suggesting that the inequality (5) may be close to being *tight* (that is, to holding with equality).

2.1 The NCC in theoretical models

The NCC is a convenient and flexible way to restrict the set of stochastic discount factors under consideration. (It would, for example, fail badly in a risk-neutral economy, that is, if M_T were deterministic.) For the NCC to hold, we need the SDF to be negatively correlated with the return R_T; this will be the case for any asset that even roughly approximates the idealized notion of 'the market' in economic models. We also need the SDF to be volatile, as is the case empirically (Hansen and Jagannathan (1991)).

The first example of this section indicates, in a conditionally lognormal setting, why the NCC is likely to hold in practice. It shows, as special cases, that the NCC holds in several leading macro-finance models. (All proofs for this section are in the appendix.)

Example 1. Suppose that the SDF M_T and return R_T are conditionally lognormal and write $r_{f,t} = \log R_{f,t}$, $\mu_{R,t} = \log \mathbb{E}_t R_T$, and $\sigma^2_{R,t} = \operatorname{var}_t \log R_T$. Then the NCC is equivalent to the assumption that the conditional Sharpe ratio of the asset, $\lambda_t \equiv (\mu_{R,t} - r_{f,t})/\sigma_{R,t}$, exceeds its conditional volatility, $\sigma_{R,t}$.

The NCC therefore holds in any conditionally lognormal model in which the market's conditional Sharpe ratio is higher than its conditional volatility. Empirically, the Sharpe ratio of the market is on the order of 50% while its volatility is on the order of 16%, so it is unsurprising that this property holds in the calibrated models of Campbell and Cochrane (1999), Bansal and Yaron (2004), Bansal, Kiku, Shaliastovich and Yaron (2012) and Campbell, Giglio, Polk and Turley (2012), among many others.

The special feature of the lognormal setting is that real-world volatility and risk-neutral volatility are one and the same thing.[4] So if an asset's Sharpe ratio is larger

[4]More precisely, $\operatorname{var}_t \log R_T = \operatorname{var}^*_t \log R_T$ if M_T and R_T are jointly lognormal under the real-world measure, conditional on time-t information.

than its (real-world or risk-neutral) volatility, then its expected excess return is larger than its (real-world or risk-neutral) variance. That is, by (4), the NCC holds.

Unfortunately, the lognormality assumption is inconsistent with well-known properties of index option prices. The most direct way to see this is to note that equity index options exhibit a volatility smile: Black–Scholes implied volatility varies across strikes, holding option maturity constant. This concern motivates the next example, which provides an interpretation of the NCC that is not dependent on a lognormality assumption.

Example 2. Suppose that there is an unconstrained investor who maximizes expected utility over next-period wealth, whose wealth is fully invested in the market, and whose relative risk aversion (which need not be constant) is at least one at all levels of wealth. Then the NCC holds for the market return.

Moreover, if (but *not* only if) the investor has log utility, the covariance term in (4) is identically zero; then, the inequality (5) holds with *equality*, and $\mathbb{E}_t R_T - R_{f,t} = \frac{1}{R_{f,t}} \operatorname{var}_t^* R_T$.

Example 2 has the attractive feature that it does not require that the identity of the investor whose wealth is fully invested in the market should be fixed over time; thus it allows for the possibility that the portfolio holdings and beliefs of (and constraints on) different investors are highly heterogeneous over time. Moreover, it does not require that all investors are fully invested in the market, that all investors are unconstrained, or that all investors are rational. Under the interpretation of Example 2, the question answered by this paper is simply this: What expected return must be perceived by an unconstrained investor with log utility who chooses to hold the market?

But, by focussing on a one-period investor, the example abstracts from intertemporal issues, and therefore from the presence of state variables that affect the value function. To the extent that we are interested in the behavior of long-lived utility-maximizing investors, we want to allow for the fact that investment opportunities vary over time, as in the framework of Merton (1973). When will the NCC hold in (a discrete-time analog of) Merton's framework? Example 1 provided one answer to this question, but we can also frame sufficient conditions directly in terms of the properties of preferences and state variables, as in the next example (in which the driving random variables are Normal, as in Example 1; this assumption will shortly be relaxed).

Example 3a. Suppose, in the notation of Cochrane (2005, pp. 166–7), that the SDF takes the form

$$M_T = \beta \frac{V_W(W_T, z_{1,T}, \ldots, z_{N,T})}{V_W(W_t, z_{1,t}, \ldots, z_{N,t})},$$

where W_T is the time-T wealth of a risk-averse investor whose wealth is fully invested in the market, so that $W_T = (W_t - C_t)R_T$ (where C_t denotes the investor's time-t consumption and R_T the return on the market); V_W is the investor's marginal value of wealth; and $z_{1,T}, \ldots, z_{N,T}$ are state variables, with signs chosen so that V_W is weakly decreasing in each (so a high value of $z_{1,T}$ is good news, just as a high value of W_T is good news). Suppose also that

(i) Risk aversion is sufficiently high: $-WV_{WW}/V_W \geq 1$ at all levels of wealth W and all values of the state variables.

(ii) The market return, R_T, and state variables, $z_{1,T}, \ldots, z_{N,T}$, are increasing functions of conditionally Normal random variables with (weakly) positive pairwise correlations.

Then the NCC holds for the market return.

Condition (i) imposes an assumption that risk aversion is at least one, as in Example 2; again, risk aversion may be wealth- and state-dependent. Condition (ii) ensures that the movements of state variables do not undo the logic of Example 1. To get a feel for it, consider a model with a single state variable, the price-dividend ratio of the market (perhaps as a proxy for the equity premium, as in Campbell and Viceira (1999)).[5] For consistency with the sign convention on the state variables, we need the marginal value of wealth to be weakly decreasing in the price-dividend ratio. It is intuitively plausible that the marginal value of wealth should indeed be high in times when valuation ratios are low; and this holds in Campbell and Viceira's setting, in the power utility case, if risk aversion is at least one.[6] Then condition (iii) amounts to the (empirically extremely

[5]The price-dividend ratio is positive, so evidently cannot be Normally distributed; this is why it is important that condition (ii) allows for state variables to be *arbitrary* increasing functions of Normal random variables. For instance, we may want to assume that the log price-dividend ratio is conditionally Normal, as Campbell and Viceira do.

[6]Campbell and Viceira also allow for Epstein–Zin preferences, which will be handled separately below.

plausible) requirement that the correlation between the wealth of the representative investor and the market price-dividend ratio is positive. Equivalently, we need the return on the market and the market price-dividend ratio to be positively correlated. Again, this holds in Campbell and Viceira's calibration.

Example 3a assumes that the investor is fully invested in the market. Roll (1977) famously criticized empirical tests of the CAPM by pointing out that stock market indices are imperfect proxies for the idealized notion of 'the market' that may not fully capture risks associated with labor or other sources of income. Without denying the force of this observation, the implicit position taken is that although the S&P 500 index is not the sum total of all wealth, it *is* reasonable to ask, as a benchmark, what equity premium would be perceived by someone fully invested in the S&P 500. (In contrast, it would be much less reasonable to assume that some investor holds all of his wealth in gold, in order to estimate the expected return on gold.)

Nonetheless, one may want to allow part of the investor's wealth to be held in assets other than the equity index. The next example generalizes Example 3a to do so. It also generalizes in another direction, by allowing the driving random variables to be non-Normal.

Example 3b. Modify Example 3a by assuming that only a fraction α_t of wealth net of consumption is invested in 'the market' (that is, in the equity index that is the focus of this paper), with the remainder invested in some other asset or portfolio of assets that earns the gross return $R_T^{(i)}$:

$$W_T = \underbrace{\alpha_t(W_t - C_t)R_T}_{\text{market wealth, } W_{M,T}} + \underbrace{(1-\alpha_t)(W_t - C_t)R_T^{(i)}}_{\text{non-market wealth}}.$$

If the signs of state variables are chosen as in Example 3a, and if

(i) Risk aversion is sufficiently high: $-WV_{WW}/V_W \geq W_T/W_{M,T}$.

(ii) $R_T, R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}$ are *associated* random variables.

then the NCC holds for the market return.

Condition (i) shows that we can allow the investor's wealth to be less than fully invested in the market (for example, in bonds, housing, and human capital), so long as he cares more about the position he does have—that is, has higher risk aversion.

If, say, at least a third of the investor's time-T wealth is attributable to the fraction invested in the market, then the NCC holds so long as risk aversion is at least three.

The concept of *associated* random variables (Esary, Proschan and Walkup (1967)) extends the concept of nonnegative correlation in a manner that can be extended to the multivariate setting. In particular, jointly Normal random variables are associated if and only if they are nonnegatively correlated (Pitt (1982)), and increasing functions of associated random variables are associated; thus Example 3a is a special case of Example 3b.

The next example handles models, such as Wachter (2013), that are neither conditionally lognormal nor feature investors with time-separable utility.

Example 4a. Suppose that there is a representative agent with Epstein–Zin (1989) preferences. If (i) risk aversion $\gamma \geq 1$ and elasticity of intertemporal substitution $\psi \geq 1$, and (ii) the market return R_T and wealth-consumption ratio W_T/C_T are associated, then the NCC holds for the market return.

As special cases, condition (ii) would hold if, say, the log return $\log R_T$ and log wealth-consumption ratio $\log W_T/C_T$ are both Normal and nonnegatively correlated; or if the elasticity of intertemporal substitution $\psi = 1$, since then the wealth-consumption ratio is constant (and hence, trivially, associated with the market return). This second case covers Wachter's (2013) model with time-varying disaster risk.

Example 4b. If there is a representative investor with Epstein–Zin (1989) preferences, with risk aversion $\gamma = 1$ and arbitrary elasticity of intertemporal substitution then the NCC holds for the market return *with equality*. This case was considered (and not rejected) by Epstein and Zin (1991) and Hansen and Jagannathan (1991).

2.2 Estimates of $\operatorname{cov}(M_T R_T, R_T)$ in linear factor models

We can also ask whether the NCC holds in linear factor models, in the style of Fama and French (1996), that aim to account for the cross-sectional variation in average stock returns. Consider a linear factor model of the SDF in the form

$$M = a_1 + a_2(R - R_f) + a_3 SMB + a_4 HML, \tag{7}$$

where the three factors are the excess returns on the market $(R - R_f)$, on 'size' (SMB), and on 'value' (HML), as in Fama and French (1996). (Below, I also consider a lin-

ear factor model including a momentum factor.) The coefficients $a_1, \ldots, a_4$ are estimated using GMM with 27 test assets: the riskless asset, the market, and 25 portfolios double-sorted on size and book-to-market. Since there are 27 moment conditions, the coefficients $a_1, \ldots, a_4$ are overidentified; I use the identity matrix to weight the moment conditions. (Appendix C reports very similar results obtained with a two-stage approach in which the weighting matrix is estimated in the first stage.) The data, which was downloaded from Kenneth French's website, is monthly, and runs from July 1926 until February 2014.

I report estimates for the full sample; for the pre-'63 sample (July 1926–December 1962); for the post-'63 sample (January 1963–February 2014); and for the post-'96 sample (January 1996–February 2014), to check that the recent time period over which I have option data is representative of the full sample.

The coefficient estimates are shown in the top half of Table 1, with standard errors in parentheses. The time-series averages of the estimated SDF are 0.995 for the full sample; 0.996 for the pre-'63 sample; 0.995 for the post-'63 sample; and 0.996 for the post-'96 sample. The time-series standard deviations of the estimated SDF are 0.128 for the full sample; 0.097 for the pre-'63 sample; 0.225 for the post-'63 sample; and 0.182 for the post-'96 sample.

The bottom half of Table 1 adds a monthly momentum factor (MOM), so the SDF is

$$M = a_1 + a_2(R - R_f) + a_3 SMB + a_4 HML + a_5 MOM. \tag{8}$$

The model is estimated as before, except that I now also include ten portfolios formed monthly on momentum using NYSE prior (2-12) return decile breakpoints, for a total of 37 test assets. Again, all data is taken from Ken French's website. Sample periods are almost identical to the above analysis: the full sample is January 1927–December 2013; the pre-'63 sample is January 1927–December 1962; the post-'63 sample is January 1963–December 2013; and the post-'96 sample is January 1996–December 2013.

The time-series averages of the estimated SDF are 0.995 for the full sample; 0.997 for the pre-'63 sample; 0.995 for the post-'63 sample; and 0.996 for the post-'96 sample. The presence of the momentum factor and portfolios increases time-series standard deviations of the estimated SDF substantially, to 0.271 for the full sample; 0.261 for the pre-'63 sample; 0.310 for the post-'63 sample; and to 0.226 for the post-'96 sample.

Based on the estimated SDFs, the rightmost column of Table 1 shows sample esti-

	constant	$R_M - R_f$	SMB	HML	—	$\widehat{\text{cov}}(M_T R_T, R_T)$
Full sample	1.013	−0.945	−0.324	−2.944	—	−0.0016
	(0.007)	(0.649)	(0.970)	(0.871)	—	(0.0017)
Jul '26–Dec '62	1.009	−0.982	−0.002	−1.090	—	−0.0020
	(0.010)	(0.942)	(1.434)	(1.433)	—	(0.0031)
Jan '63–Feb '14	1.045	−2.874	−2.587	−7.413	—	−0.0019
	(0.018)	(1.137)	(1.472)	(1.607)	—	(0.0020)
Jan '96–Feb '14	1.028	−1.984	−3.000	−4.948	—	−0.0015
	(0.026)	(1.735)	(2.207)	(2.350)	—	(0.0034)
	constant	$R_M - R_f$	SMB	HML	MOM	$\widehat{\text{cov}}(M_T R_T, R_T)$
Full sample	1.072	−2.375	−0.648	−5.489	−5.572	−0.0018
	(0.020)	(0.746)	(1.011)	(1.131)	(1.033)	(0.0020)
Jan '27–Dec '62	1.071	−2.355	−0.587	−3.882	−5.552	−0.0021
	(0.029)	(1.034)	(1.747)	(2.163)	(1.565)	(0.0041)
Jan '63–Dec '13	1.092	−3.922	−2.400	−9.020	−5.152	−0.0020
	(0.029)	(1.272)	(1.475)	(1.795)	(1.427)	(0.0022)
Jan '96–Dec '13	1.047	−3.231	−2.327	−5.789	−2.548	−0.0017
	(0.034)	(1.981)	(2.224)	(2.491)	(1.637)	(0.0036)

Table 1: Estimates of coefficients in the factor models (7) and (8), and of $\text{cov}(M_T R_T, R_T)$. Standard errors are in brackets.

mates of unconditional covariance, $\text{cov}(M_T R_T, R_T)$,[7] in each sample period and with and without momentum; standard errors are in parentheses. The estimates are extremely stable across different sample periods and hardly change when the momentum factor and portfolios are included. Consistent with the NCC, the estimates are negative in every sample period and in both tables. They are also close to zero in economic terms, and not significantly different from zero in statistical terms, suggesting that the inequality (5) may be close to being tight.

3 Risk-neutral variance and the SVIX index

We now turn to the question of measuring the risk-neutral variance that appears on the right-hand side of (5). The punchline will be that risk-neutral variance is uniquely pinned down by European option prices, by a static no-arbitrage argument. To streamline the exposition, I will temporarily assume that the prices of European call and put options expiring at time T on the asset with return R_T are perfectly observable at all strikes K; this unrealistic assumption will be relaxed below.

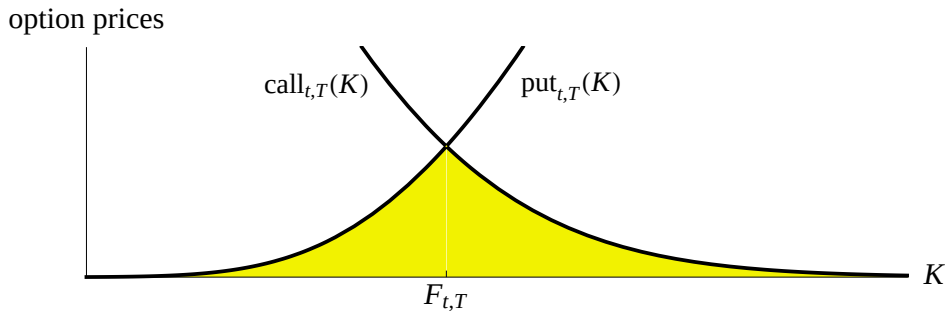


Figure 1: The prices, at time t, of call and put options expiring at time T.

Figure 1 plots a generic collection of time-t prices of calls expiring at time T with strike K (written $\text{call}_{t,T}(K)$) and of puts expiring at time T with strike K (written $\text{put}_{t,T}(K)$). The figure illustrates two well-known facts that will be useful. First, call and put prices are convex functions of strike, K. (Any non-convexity would provide a static arbitrage opportunity.) This property will allow us, below, to deal with the issue that option prices are only observable at a limited set of strikes. Second, the forward

[7] This is the unconditional expectation of $\text{cov}_t(M_T R_T, R_T)$, because $\mathbb{E}\left[\text{cov}_t(M_T R_T, R_T)\right] = \mathbb{E}\left[\mathbb{E}_t(M_T R_T^2) - \mathbb{E}_t(M_T R_T)\,\mathbb{E}_t\, R_T\right] = \mathbb{E}(M_T R_T^2) - \mathbb{E}(M_T R_T)\,\mathbb{E}\, R_T = \text{cov}(M_T R_T, R_T)$.

price of the underlying asset, $F_{t,T}$, which satisfies

$$F_{t,T} = \mathbb{E}_t^* S_T, \tag{9}$$

can be determined by observing the strike at which call and put prices are equal, i.e., $F_{t,T}$ is the unique solution x of the equation $\text{call}_{t,T}(x) = \text{put}_{t,T}(x)$. This fact follows from put-call parity; it means that the forward price can be backed out from time-t option prices.

We want to measure $\frac{1}{R_{f,t}} \text{var}_t^* R_T$. I assume that the dividends earned between times t and T are known at time t and paid at time T,[8] so that

$$\frac{1}{R_{f,t}} \text{var}_t^* R_T = \frac{1}{S_t^2} \left[\frac{1}{R_{f,t}} \mathbb{E}_t^* S_T^2 - \frac{1}{R_{f,t}} \left(\mathbb{E}_t^* S_T \right)^2 \right]. \tag{10}$$

We can deal with the second term inside the square brackets using equation (9), so the challenge is to calculate $\frac{1}{R_{f,t}} \mathbb{E}_t^* S_T^2$. This is the price of the 'squared contract'—that is, the price of a claim to S_T^2 paid at time T.

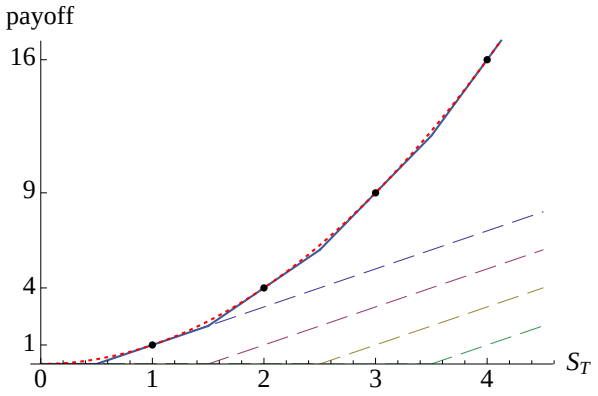


Figure 2: The payoff S_T^2 (dotted line); and the payoff on a portfolio of options (solid line), consisting of two calls with strike $K = 0.5$, two calls with $K = 1.5$, two calls with $K = 2.5$, two calls with $K = 3.5$, and so on. Individual option payoffs are indicated by dashed lines.

How can we price this contract, given put and call prices as illustrated in Figure 1? Suppose we buy two call options with a strike of $K = 0.5$; two calls with a strike of

[8] If dividends are not known ahead of time, it is enough to assume that prices and dividends are (weakly) positively correlated, since then $\text{var}_t^* R_T \geq \text{var}_t^*(S_T/S_t)$, so that using $\frac{1}{R_{f,t}} \text{var}_t^*(S_T/S_t)$ instead of the ideal lower bound, $\frac{1}{R_{f,t}} \text{var}_t^* R_T$, is conservative.

$K = 1.5$; two calls with a strike of $K = 2.5$; two calls with a strike of $K = 3.5$; and so on, up to arbitrarily high strikes. The payoffs on the individual options are shown as dashed lines in Figure 2, and the payoff on the portfolio of options is shown as a solid line. The idealized payoff S_T^2 is shown as a dotted line. The solid and dotted lines almost perfectly overlap, illustrating that the payoff on the portfolio is almost exactly S_T^2 (and it is *exactly* S_T^2 at integer values of S_T). Therefore, the price of the squared contract is approximately the price of the portfolio of options:

$$\frac{1}{R_{f,t}}\,\mathbb{E}_t^*\,S_T^2 \approx 2\sum_{K=0.5,1.5,\ldots} \text{call}_{t,T}(K). \tag{11}$$

I show in the appendix that the squared contract can be priced exactly by replacing the sum with an integral:

$$\frac{1}{R_{f,t}}\,\mathbb{E}_t^*\,S_T^2 = 2\int_{K=0}^{\infty} \text{call}_{t,T}(K)\,dK. \tag{12}$$

In practice, of course, option prices are not observable at *all* strikes K, so we will need to approximate the idealized integral (12) by a sum along the lines of (11). To see how this will affect the results, notice that Figure 2 also demonstrates a subtler point: the option portfolio payoff is not just equal to the 'squared payoff' at integers, it is *tangent* to it, so that the payoff on the portfolio of options very closely approximates *and is always less than or equal to* the ideal squared payoff. As a result, the sum over call prices in (11) will be slightly *less* than the integral over call prices in (12). This implies that the bounds presented are robust to the fact that option prices are not observable at all strikes: they would be *even higher* if all strikes were observable. Section 4.1 expands on this point.

Finally, since deep-in-the-money call options are neither liquid in practice nor intuitive to think about, it is convenient to split the range of integration into two and use put-call parity to replace in-the-money call prices with out-of-the-money put prices. Doing so, and substituting the result back into (10), we find that

$$\frac{1}{R_{f,t}}\,\text{var}_t^*\,R_T = \frac{2}{S_t^2}\left[\int_0^{F_{t,T}} \text{put}_{t,T}(K)\,dK + \int_{F_{t,T}}^{\infty} \text{call}_{t,T}(K)\,dK\right]. \tag{13}$$

The expression in the square brackets is the shaded area shown in Figure 1.

The right-hand side of (13) is strongly reminiscent of the definition of the VIX index. To bring out the connection it will be helpful to define an index, SVIX_t, via the

formula

$$\mathrm{SVIX}_t^2 = \frac{2}{(T-t)\cdot R_{f,t}\cdot S_t^2}\left[\int_0^{F_{t,T}} \mathrm{put}_{t,T}(K)\,dK + \int_{F_{t,T}}^{\infty} \mathrm{call}_{t,T}(K)\,dK\right]. \tag{14}$$

The SVIX index measures the annualized risk-neutral variance of the realized excess return: comparing equations (13) and (14), we see that

$$\mathrm{SVIX}_t^2 = \frac{1}{T-t}\,\mathrm{var}_t^*(R_T/R_{f,t}). \tag{15}$$

4 A lower bound on the equity premium

We can summarize the results of previous sections by inserting (13) into inequality (5). This gives a lower bound on the expected excess return of any asset that obeys the NCC:

$$\mathbb{E}_t\,R_T - R_{f,t} \geq \frac{2}{S_t^2}\left[\int_0^{F_{t,T}} \mathrm{put}_{t,T}(K)\,dK + \int_{F_{t,T}}^{\infty} \mathrm{call}_{t,T}(K)\,dK\right] \tag{16}$$

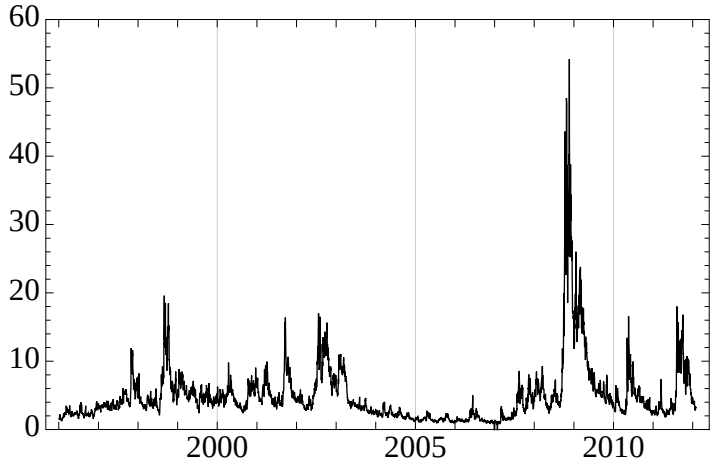
or, in terms of the SVIX index,

$$\frac{1}{T-t}\left(\mathbb{E}_t\,R_T - R_{f,t}\right) \geq R_{f,t}\cdot \mathrm{SVIX}_t^2. \tag{17}$$

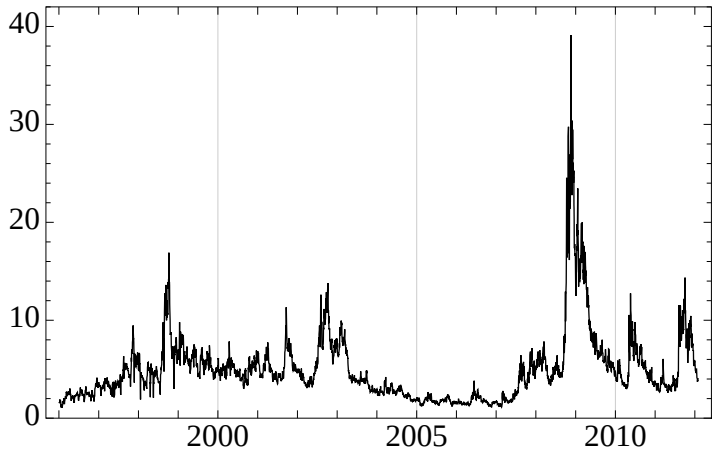
The bound will be applied in the case of the S&P 500; from now on, R_T always refers to the gross return on the S&P 500 index. I construct a time series of the lower bound from January 4, 1996 to January 31, 2012 using option price data from *OptionMetrics*; Appendix B.1 contains full details of the procedure. I compute the bound for time horizons $T-t = 1$, 2, 3, 6, and 12 months. I report results in annualized terms; that is, both sides of the above inequality are multiplied by $\frac{1}{T-t}$ with t and T measured in years (so, for example, monthly expected returns are multiplied by 12 to convert them into annualized terms).

Figure 3a plots the lower bound, annualized and in percentage points, at the 1-month horizon. Figures 3b and 3c repeat the exercise at 3-month and 1-year horizons. Table 2 reports the mean, standard deviation, and various quantiles of the distribution of the lower bound in the daily data for horizons between 1 month and 1 year.

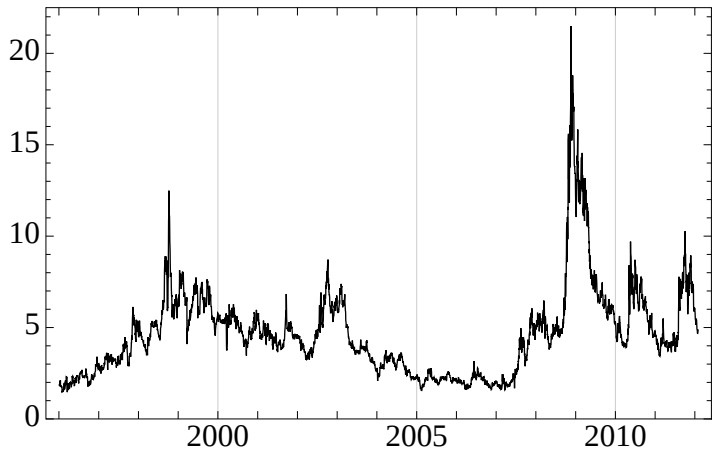
The mean of the lower bound over the whole sample is 5.00% at the monthly horizon. This number is strikingly close to typical estimates of the unconditional equity premium, which supports the suggestion made in Section 2.2 that the bound



(a) 1 month



(b) 3 month



(c) 1 year

Figure 3: The lower bound on the annualized equity premium at different horizons.

horizon	mean	s.d.	min	1%	10%	25%	50%	75%	90%	99%	max
1 mo	5.00	4.60	0.83	1.03	1.54	2.44	3.91	5.74	8.98	25.7	55.0
2 mo	5.00	3.99	1.01	1.20	1.65	2.61	4.11	5.91	8.54	23.5	46.1
3 mo	4.96	3.60	1.07	1.29	1.75	2.69	4.24	5.95	8.17	21.4	39.1
6 mo	4.89	2.97	1.30	1.53	1.95	2.88	4.39	6.00	7.69	16.9	29.0
1 yr	4.64	2.43	1.47	1.64	2.07	2.81	4.35	5.72	7.19	13.9	21.5

Table 2: Mean, standard deviation, and quantiles of the lower bound on the equity premium, $R_{f,t} \cdot \text{SVIX}_t^2$, at various horizons (annualized and measured in %).

may be fairly tight: that is, it seems that the inequality (16) may approximately hold with equality. Below, I provide further tests of this possibility and develop some of its implications.

The time-series average of the lower bound is lower at the annual horizon than it is at the monthly horizon where the data quality is best (perhaps because of the existence of trades related to VIX, which is itself a monthly index). It is likely that this reflects a less liquid market in 1-year options, with a smaller range of strikes traded, rather than an interesting economic phenomenon. I discuss this further in Section 4.1 below.

The lower bound is volatile and right-skewed. At the annual horizon the equity premium varies from a minimum of 1.22% to a maximum of 21.5% over my sample period. But variation at the one-year horizon masks even more dramatic variation over shorter horizons. The monthly lower bound averaged only 1.86% (annualized) during the "Great Moderation" years 2004–2006, but peaked at 55.0%—more than 10 standard deviations above the mean—in November 2008, at the height of the subprime crisis. Indeed, the lower bound hit peaks at all horizons during the recent crisis, notably from late 2008 to early 2009 as the credit crisis gathered steam and the stock market fell, but also around May 2010, coinciding with the beginning of the European sovereign debt crisis. Other peaks occur during the LTCM crisis in late 1998; during the days following September 11, 2001; and during a period in late 2002 when the stock market was hitting new lows following the end of the dotcom boom.

Figure 11, in the appendix, shows that there was an increase in daily volume and open interest in S&P 500 index options over my sample period. The peaks in SVIX in 2008, 2010, and 2011 are associated with spikes in volume.

4.1 Robustness of the lower bound

Were option markets illiquid during the subprime crisis? One potential concern is that option markets may have been illiquid during periods of extreme stress. If so, one would expect to see a significant disparity between bounds based on mid-market option prices, such as those shown in Figure 3, and bounds based on bid or offer prices, particularly in periods such as November 2008. Thus it is possible in principle that the lower bounds would decrease significantly if bid prices were used. Figure 12, in the appendix, plots bounds calculated from bid prices. Reassuringly, the results are very similar: the lower bound is high at all horizons whether mid or bid prices are used.

Option prices are only observable at a discrete range of strikes. Two issues arise when implementing the lower bound. Fortunately, both issues mean that the numbers presented in this paper are conservative: with ideal data, the lower bound would be even higher.

First, we do not observe option prices at all strikes K between 0 and ∞. This means that the range of integration in the integral we would ideally like to compute—the shaded area in Figure 1—is truncated. Obviously, this will cause us to underestimate the integral in practice. This effect is likely to be strongest at the 1-year horizon, because (in my dataset) 1-year options are less liquid than shorter-dated options.

Second, even within the range of observable strikes, prices are only available at a discrete set of strikes. Thus the idealized lower bound that emerges from the theory in the form of an integral (over option prices at all strikes) must be approximated by a sum (over option prices at observable strikes). What effect will this have? In the discussion of Figure 1, I provided an example in which the price of a particular portfolio of calls with a discrete set of strikes would very slightly underestimate the idealized measure, and hence be conservative. The general case, using out-of-the-money puts and calls, is handled in Appendix B.2. The conclusion is the same: discretization leads to underestimates of risk-neutral variance, and hence to a conservative bound.

5 Might the lower bound be tight?

Two facts make it reasonable to ask whether the lower bound might in fact be tight. First, the results of Section 2.2 yielded estimates of the covariance term $\text{cov}(M_T R_T, R_T)$ that are statistically and economically close to zero. Second, the time-series average

of the lower bound in recent data is approximately 5% in annualized terms, a number close to conventional estimates of the equity premium. Over the period 1951–2000, Fama and French (2002) estimate the unconditional average equity premium to be 3.83% or 4.78%, based on dividend and earnings growth respectively.[9]

We therefore want to test the hypothesis that $\frac{1}{T-t}\left(\mathbb{E}_t R_T - R_{f,t}\right) = R_{f,t} \cdot \mathrm{SVIX}_t^2$. Table 3 shows the results of regressions

$$\frac{1}{T-t}\left(R_T - R_{f,t}\right) = \alpha + \beta \times R_{f,t} \cdot \mathrm{SVIX}_t^2 + \varepsilon_T\,, \tag{18}$$

together with robust Hansen–Hodrick standard errors that account for heteroskedasticity and overlapping observations. The null hypothesis that $\alpha = 0$ and $\beta = 1$ is not rejected at any horizon. The point estimates on β are close to 1 at all horizons, lending further support to the possibility that the lower bound is tight. This is encouraging because, as Goyal and Welch (2008) emphasize, this period is one in which conventional predictive regressions fare poorly.

horizon	$\widehat{\alpha}$	s.e.	$\widehat{\beta}$	s.e.	R^2	R^2_{OS}
1 mo	0.012	[0.064]	0.779	[1.386]	0.34%	0.42%
2 mo	−0.002	[0.068]	0.993	[1.458]	0.86%	1.11%
3 mo	−0.003	[0.075]	1.013	[1.631]	1.10%	1.49%
6 mo	−0.056	[0.058]	2.104	[0.855]	5.72%	4.86%
1 yr	−0.029	[0.093]	1.665	[1.263]	4.20%	4.73%

Table 3: Coefficient estimates for the regression (18).

One might worry that these results are entirely driven by the period in 2008 and 2009 in which volatility spiked and the stock market crashed before recovering strongly. To address this concern, Table 6, in the appendix, shows the result of deleting all observations that overlap with the period August 1, 2008–July 31, 2009. Over horizons of 1, 2, and 3 months, deleting this period in fact *increases* the forecastability of returns by SVIX, reflecting the fact that the market continued to drop for a time after volatility spiked up in November 2008. On the other hand, the subsequent strong

[9]These are the 'bias-adjusted' figures presented in their Table IV. In an interview with Richard Roll available on the AFA website at http://www.afajof.org/details/video/2870921/Eugene-Fama-Interview.html, Fama says, "I always think of the number, the equity premium, as five per cent."

recovery of the market means that this was a period in which 1-year options successfully predicted 1-year returns, so by removing the crash from the sample, the forecasting power deteriorates at the 1-year horizon.

We now have seen from various different angles that the lower bound (16) appears to be approximately tight: (i) as shown in Table 2 and Figure 3, the average level of the lower bound over my sample is close to conventional estimates of the average equity premium; (ii) Table 3 shows that the null hypothesis that $\alpha = 0$ and $\beta = 1$ in the forecasting regression (18) is not rejected at any horizon; and (iii) the estimates of $\text{cov}(M_T R_T, R_T)$ shown in Table 1 are statistically and economically close to zero.

These observations suggest that SVIX can be used as a measure of the equity premium without estimating any parameters—that is, imposing $\alpha = 0$, $\beta = 1$ in (18), so that

$$\frac{1}{T-t}\left(\mathbb{E}_t\, R_T - R_{f,t}\right) = R_{f,t} \cdot \text{SVIX}_t^2. \tag{19}$$

To assess the performance of the forecast (19), I follow Goyal and Welch (2008) in computing an out-of-sample R-squared measure

$$R_{OS}^2 = 1 - \frac{\sum \varepsilon_t^2}{\sum \nu_t^2}, \tag{20}$$

where ε_t is the error when SVIX (more precisely, $R_{f,t} \cdot \text{SVIX}_t^2$) is used to forecast the equity premium and ν_t is the error when the historical mean equity premium (computed on a rolling basis) is used to forecast the equity premium.[10]

The rightmost column of Table 3 reports the values of R_{OS}^2 at each horizon. These out-of-sample R_{OS}^2 values can be compared with corresponding numbers for forecasts based on valuation ratios, which are the subject of a vast literature.[11] Goyal and Welch (2008) consider return predictions in the form

$$\text{equity premium}_t = a_1 + a_2 \times \text{predictor variable}_t\,, \tag{21}$$

where a_1 and a_2 are constants estimated from the data, and argue that while conventional predictor variables perform reasonably well in-sample, they perform worse out-of-sample than the rolling mean. Over their full sample (which runs from 1871 to

[10] More detail on the construction of the rolling mean is provided in the appendix.

[11] Among many others, Campbell and Shiller (1988), Fama and French (1988), Lettau and Ludvigson (2001), and Cochrane (2008) make the case for predictability. Other authors, including Ang and Bekaert (2007), make the case against.

2005, with the first 20 years used to initialize estimates of a_1 and a_2, so that predictions start in 1891), the dividend-price ratio, dividend yield, earnings-price ratio, and book-to-market ratio have negative out-of-sample R^2s of -2.06%, -1.93%, -1.78% and -1.72%, respectively. The performance of these predictors is particularly poor over Goyal and Welch's 'recent sample' (1976 to 2005), with R^2s of -15.14%, -20.79%, -5.98% and -29.31%, respectively.[12]

Campbell and Thompson (2008) confirm Goyal and Welch's finding, and respond by suggesting that the coefficients a_1 and a_2 be fixed based on *a priori* considerations. Motivated by the Gordon growth model $D/P = R - G$ (where D/P is the dividend-price ratio, R the expected return, and G expected dividend growth), Campbell and Thompson suggest making forecasts of the form

$$\text{equity premium}_t = \text{dividend-price ratio}_t + \text{dividend growth}_t - \text{real interest rate}_t$$

or, more generally,

$$\text{equity premium}_t = \text{valuation ratio}_t + \text{dividend growth}_t - \text{real interest rate}_t, \quad (22)$$

where in addition to the dividend-price ratio, Campbell and Thompson also consider earnings yields, smoothed earnings yields, and book-to-market as valuation ratios.[13] Since these forecasts are drawn directly from the data without requiring estimation of coefficients, they are a natural point of comparison for the forecast (19) suggested in this paper.

Over the full sample, the out-of-sample R^2s corresponding to the forecasts (22) range from 0.24% (using book-to-market as the valuation ratio) to 0.52% (using smoothed earnings yield) in monthly data; and from 1.85% (earnings yield) to 3.22% (smoothed earnings yield) in annual data.[14] The results are worse over Campbell and Thompson's most recent subsample, from 1980–2005: in monthly data, R^2 ranges from -0.27% (book-to-market) to 0.03% (earnings yield). In annual data, the forecasts do even more poorly, each underperforming the historical mean, with R^2s ranging from -6.20% (book-to-market) to -0.47% (smoothed earnings yield).

[12]Goyal and Welch show that the performance of an out-of-sample version of Lettau and Ludvigson's (2001) *cay* variable is similarly poor, with $\widetilde{R}^2$ of -4.33% over the full sample and -12.39% over the recent sample.

[13]The *real* interest rate is subtracted because dividend growth is measured in real terms.

[14]Out-of-sample forecasts are from 1927 to 2005, or 1956 to 2005 when book-to-market is used.

In relative terms, therefore, the out-of-sample R-squareds shown in Table 3 compare very favorably with the corresponding R-squareds for predictions based on valuation ratios. But are they too small to be interesting in absolute terms? No. Ross (2005, pp. 54–57) and Campbell and Thompson (2008) point out that high R^2 statistics in predictive regressions translate into high attainable Sharpe ratios, for the simple reason that the predictions can be used to formulate a market-timing trading strategy; and if the predictions are very good, the strategy will perform extremely well. If Sharpe ratios above some level are 'too good to be true,' then one should not expect to see R^2s from predictive regressions above some upper limit.

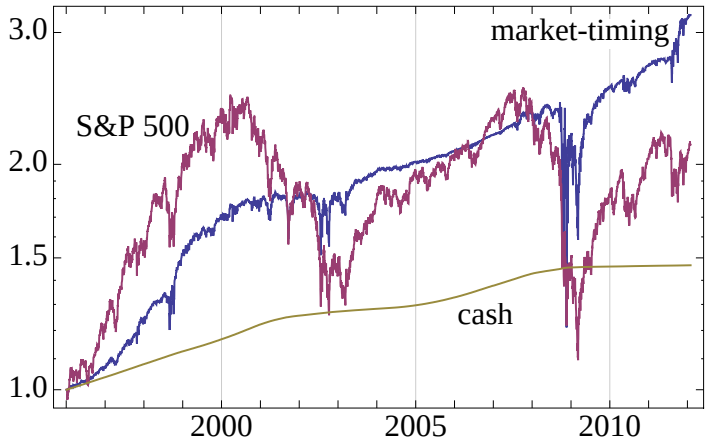


Figure 4: Cumulative returns on \$1 invested in cash, in the S&P 500 index, or in a market-timing strategy whose allocation to the market is proportional to $R_{f,t} \cdot \text{SVIX}_t^2$. Log scale.

With this thought in mind, consider using risk-neutral variance in a contrarian market-timing strategy: invest, each day, a fraction α_t in the S&P 500 index and the remaining fraction $1-\alpha_t$ at the riskless rate, where α_t is chosen proportional to 1-month SVIX^2 (scaled by the riskless rate, as on the right-hand side of (17)). The constant of proportionality has no effect on the strategy's Sharpe ratio, so I choose it such that the market-timing strategy's mean portfolio weight in the S&P 500 is 35%, with the remaining 65% in cash; the resulting median portfolio weight is 27% in the S&P 500, with 73% in cash. Figure 4 plots the cumulative return on an initial investment of \$1 in this market-timing strategy and, for comparison, on strategies that invest in the short-term interest rate or in the S&P 500 index. In my sample period, the daily

Sharpe ratio of the market is 1.35%, while the daily Sharpe ratio of the market-timing strategy is 1.97%; in other words, the out-of-sample R^2 of 0.42% reported in Table 3 is enough to deliver a 45% increase in Sharpe ratio for the market-timing strategy relative to the market itself. This exercise also illustrates an attractive feature of risk-neutral variance as a predictive variable: since it is an asset price—specifically, the price of a portfolio of options—it can be computed in daily data, or at even higher frequency, and so permits high-frequency market-timing strategies to be considered.

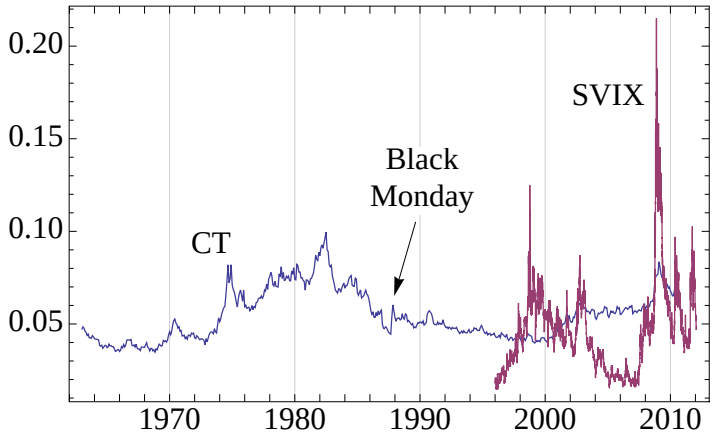


Figure 5: Equity premium forecasts based on Campbell–Thompson (2008) and on SVIX. Annual horizon.

Valuation ratios and SVIX tell qualitatively very different stories about the equity premium. Figure 5 plots the 1-year forecast $R_{f,t} \cdot \text{SVIX}_t^2$ on the same axes as the Campbell–Thompson smoothed earnings yield predictor.[15] The figure makes two things clear. First, option prices point toward a far more volatile equity premium than do valuation ratios. Second, SVIX is much less persistent than are valuation ratios, and so the SVIX predictor variable is less subject to Stambaugh (1999) bias.

It is noteworthy that SVIX forecasts a relatively high equity premium in the late 1990s. In this respect it diverges sharply from valuation-ratio-based forecasts—which then predicted a low or even negative 1-year equity premium—but, intriguingly, lines up fairly well with the expected returns reported in Gallup surveys at that time (Greenwood and Shleifer (2014)). (In Section 7, I will show that although expected returns

[15]I thank John Campbell for sharing an updated version of the dataset used in Campbell and Thompson (2008).

were high in the late 1990s, the expected return *conditional on a market downturn* was unusually low.)

But perhaps the most striking aspect of Figure 5 is the behavior of the Campbell–Thompson predictor variable on Black Monday, October 19, 1987. This was by far the worst day in stock market history: the S&P 500 index dropped by over 20%—more than twice as far as on the second-worst day in history—and yet the valuation-ratio approach suggests that the equity premium barely responded. In sharp contrast, option prices are well known to have exploded on Black Monday, implying that the equity premium was even higher than the peaks attained in November 2008.[16]

The general point that subjectively expected returns may have been very high at short horizons following Black Monday is broadly consistent with the findings of Shiller (1987), who reports the results of investor surveys that were sent out in the immediate aftermath of the crash. Although the survey questions Shiller asked are hard to compare directly with the results of the present paper, he documents that a substantial fraction of investors expected a market rebound from the crash. Shiller also reports that some investors had more nuanced expectations of market returns: for instance, some thought that the market would perform better over shorter horizons than over long horizons. The next sections explore this possibility further.

In the remainder of the paper, I treat the lower bound as tight, and measure the equity premium via the SVIX index, as in equation (19).

6 Expected returns in the long and the short run

Campbell and Shiller (1988) showed that any dividend-paying asset satisfies an (approximate) identity

$$d_t - p_t = \text{constant} + \mathbb{E}_t \sum_{j=0}^{\infty} \rho^j \left(r_{t+1+j} - \Delta d_{t+1+j} \right)$$

that relates its log dividend yield $d_t - p_t$ to expectations of future log returns r_{t+1+j} and future log dividend growth Δd_{t+1+j}. Empirically, dividend growth is approximately

[16]Figure 13, in the appendix, shows how the VXO index—that is, 1-month at-the-money implied volatility on the S&P 100—exploded on October 19, 1987. (The VIX index itself did not exist at that time.) As it turned out, the annualized return on the S&P 500 index was 81.2% over the month, and 23.2% over the year, following Black Monday.

unforecastable; to the extent that this is the case, we can absorb the terms $\mathbb{E}_t \, \Delta d_{t+1+j}$ into the constant, giving

$$d_t - p_t = \text{constant} + \mathbb{E}_t \sum_{j=0}^{\infty} \rho^j r_{t+1+j}. \tag{23}$$

Equation (23) provides intuition for why dividend yields should be informative about expected returns: treating dividend growth as approximately unforecastable, we can think of dividend yield as providing a rough measure of expected returns *over the very long run*. (A caveat: equation (23) features expected log returns rather than log expected returns. Since $\mathbb{E}_t \, r_{t+1+j} = \log \mathbb{E}_t \, R_{t+1+j} - \frac{1}{2} \operatorname{var}_t r_{t+1+j} - \sum_{n=3}^{\infty} \frac{\kappa_t^{(n)}(r_{t+1+j})}{n!}$, where $\kappa_t^{(n)}(r_{t+1+j})$ is the nth conditional cumulant of r_{t+1+j}, the gap between the two depends on the cumulants of log returns. So even if dividend growth is unforecastable, a low dividend yield may be associated with *high* expected arithmetic returns if log returns happen to be unusually volatile, right-skewed, or fat-tailed.)

In contrast, the SVIX index measures expected returns over the short run. The gap between the two is therefore informative about the gap between long-run and short-run expected returns. In the late 1990s, for example, $d_t - p_t$ was extremely low, indicating—subject to the above caveat—very low expected long-run returns (Shiller (2000)); but Figures 3a–3c show that SVIX, and hence expected *short-run* returns, were relatively *high* at that time.

We can also compare expected returns across shorter horizons. For example, Figures 3a–3c seem to show that an unusually large fraction of the elevated 1-year equity premium available in late 2008 was expected to materialize over the first few months of the 12-month period. To analyze this more formally, define the annualized *forward equity premium from T_1 to T_2* (which is known at time t) by the formula

$$\text{EP}_{T_1 \to T_2} \equiv \frac{1}{T_2 - T_1} \left(\log \frac{\mathbb{E}_t \, R_{t \to T_2}}{R_{f,t \to T_2}} - \log \frac{\mathbb{E}_t \, R_{t \to T_1}}{R_{f,t \to T_1}} \right), \tag{24}$$

and the corresponding ('spot') equity premium from time t to time T by

$$\text{EP}_{t \to T} \equiv \frac{1}{T - t} \log \frac{\mathbb{E}_t \, R_{t \to T}}{R_{f,t \to T}}.$$

Using (19) to substitute out for $\mathbb{E}_t \, R_{t \to T_1}$ and $\mathbb{E}_t \, R_{t \to T_2}$ in (24), we can write

$$\text{EP}_{T_1 \to T_2} = \frac{1}{T_2 - T_1} \log \frac{1 + \text{SVIX}^2_{t \to T_2}(T_2 - t)}{1 + \text{SVIX}^2_{t \to T_1}(T_1 - t)} \quad \text{and} \quad \text{EP}_{t \to T} = \frac{1}{T - t} \log \left(1 + \text{SVIX}^2_{t \to T}(T - t)\right).$$

(I have slightly modified previous notation to accommodate the extra time dimension; for example, $R_{t\to T_2}$ is the simple return on the market from time t to time T_2, $R_{f,t\to T_1}$ is the riskless return from time t to time T_1, and $\text{SVIX}^2_{t\to T_2}$ is the time-t level of the SVIX index calculated using options expiring at T_2.)

The definition (24) is chosen so that, for arbitrary $T_1, \ldots, T_N$, we have the decomposition

$$\text{EP}_{t\to T_N} = \frac{T_1 - t}{T_N - t}\text{EP}_{t\to T_1} + \frac{T_2 - T_1}{T_N - t}\text{EP}_{T_1\to T_2} + \cdots + \frac{T_N - T_{N-1}}{T_N - t}\text{EP}_{T_{N-1}\to T_N} \qquad (25)$$

which expresses the long-horizon equity premium $\text{EP}_{t\to T_N}$ as a weighted average of forward equity premia, exactly analogous to the relationship between spot and forward bond yields.

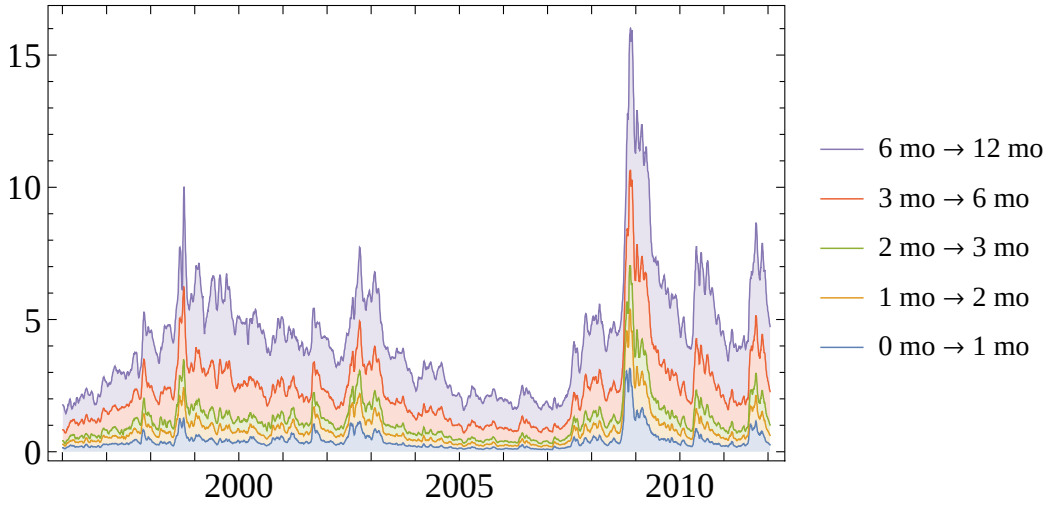


Figure 6: The term structure of equity premia. 10-day moving average.

Figure 6 shows how the annual equity premium previously plotted in Figure 3c decomposes into a one-month spot premium plus forward premia from one to two, two to three, three to six, and six to twelve months. The figure stacks the *un*annualized forward premia—terms of the form $(T_n - T_{n-1})/(T_N - t)\text{EP}_{T_{n-1}\to T_n}$—which add up to the annual equity premium, as shown in (25). For example, on any given date t, the gap between the top two lines represents the contribution of the unannualized 6-month-6-month-forward equity premium, $\frac{1}{2}\text{EP}_{t+6\text{mo}\to t+12\text{mo}}$, to the annual equity premium, $\text{EP}_{t\to t+12\text{mo}}$.

The figure reveals that in 'normal' times, the unannualized 6-month-6-month-forward equity premium contributes approximately half of the annual equity premium,

as might have been expected. But more interestingly, it also shows that at times of stress, much of the annual equity premium is compressed into the first few months. For example, in November 2008 when the annual equity premium reached its peak over this sample period, about a third of the expected equity premium over the entire year from November 2008 to November 2009 can be attributed to the expected (unannualized) equity premium over the two months from November 2008 to January 2009.

7 The up-premium and the down-premium

Previous sections have argued for SVIX^2 as a measure of the equity premium, i.e. that

$$\mathbb{E}_t\, R_T - R_{f,t} = \frac{2}{S_t^2}\left[\int_0^{F_{t,T}} \text{put}_{t,T}(K)\, dK + \int_{F_{t,T}}^{\infty} \text{call}_{t,T}(K)\, dK\right].$$

It turns out that the two components of SVIX^2,

$$\frac{2}{S_t^2}\int_0^{F_{t,T}} \text{put}_{t,T}(K)\, dK \quad \text{and} \quad \frac{2}{S_t^2}\int_{F_{t,T}}^{\infty} \text{call}_{t,T}(K)\, dK,$$

which I will call down-SVIX^2 and up-SVIX^2 respectively, can each be given a nice interpretation.

To do so, it will be convenient to think from the perspective of an investor with log utility who chooses to hold the market. We have already seen in Example 2 of Section 2.1 that the inequality (5) holds with equality in this case:[17] writing $\widetilde{\mathbb{E}}$ to emphasize that we are inferring the subjective expectations of the log investor, we have $\widetilde{\mathbb{E}}\, R_T - R_{f,t} = \frac{1}{R_{f,t}} \text{var}_t^*\, R_T$. Such an investor must perceive the market as growth-optimal, so $M_T = 1/R_T$ is a stochastic discount factor with respect to the log investor's beliefs. Thus, for any tradable time-T payoff X_T,

$$\text{time-}t \text{ price of a claim to } X_T = \frac{1}{R_{f,t}}\, \mathbb{E}_t^*\, X_T = \widetilde{\mathbb{E}}_t\, \frac{X_T}{R_T}.$$

Applying this with $X_T = Y_T R_T$—where Y_T is some random variable of interest—we can compute the log investor's subjective expectation of Y_T as

$$\widetilde{\mathbb{E}}_t\, Y_T = \frac{1}{R_{f,t}}\, \mathbb{E}_t^*\, (Y_T R_T) = \text{price of a claim to } Y_T R_T. \tag{26}$$

[17] Note, however, that although sufficient, it is not *necessary* for there to be a log investor for the covariance term to equal zero, and hence for the lower bound to be tight.

Equation (26) will be useful if the claim $Y_T R_T$ can be replicated and hence priced.

For example, if $Y_T = \mathbf{1}\{R_T > R_{f,t}\}$ then the payoff $Y_T R_T$ can be replicated by buying an at-the-money-forward digital (or binary) call and an at-the-money-forward call option. More precisely, we have[18]

$$\widetilde{\mathbb{P}}_t\left(R_T > R_{f,t}\right) = \widetilde{\mathbb{E}}_t\left(\mathbf{1}\left\{R_T > R_{f,t}\right\}\right) = \underbrace{-R_{f,t}\,\mathrm{call}'_{t,T}(F_{t,T})}_{\mathbb{P}^*_t\left(R_T > R_{f,t}\right)} + \frac{\mathrm{call}_{t,T}(F_{t,T})}{S_t}. \tag{27}$$

Equation (27) shows that the log investor's perceived probability of an up-move exceeds the risk-neutral probability by an amount, $\mathrm{call}_{t,T}(F_{t,T})/S_t$, that depends on the price of an at-the-money-forward option.

If we set $Y_T = (R_T - R_{f,t})\mathbf{1}\{R_T > R_{f,t}\}$ and $Y_T = (R_T - R_{f,t})\mathbf{1}\{R_T < R_{f,t}\}$ in (26), we find that

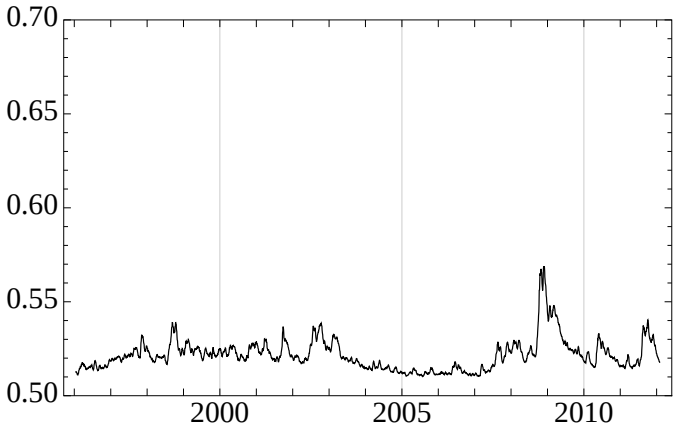
$$\frac{\widetilde{\mathbb{E}}_t\left[(R_T - R_{f,t})\mathbf{1}\left\{R_T > R_{f,t}\right\}\right]}{T-t} = \underbrace{\frac{R_{f,t}\,\mathrm{call}_{t,T}(F_{t,T})}{(T-t)S_t}}_{\frac{\mathbb{E}^*_t\left[(R_T - R_{f,t})\mathbf{1}\left\{R_T > R_{f,t}\right\}\right]}{T-t}} + R_{f,t}\,\text{up-SVIX}^2_t \tag{28}$$

$$\frac{\widetilde{\mathbb{E}}_t\left[(R_T - R_{f,t})\mathbf{1}\left\{R_T < R_{f,t}\right\}\right]}{T-t} = \underbrace{-\frac{R_{f,t}\,\mathrm{call}_{t,T}(F_{t,T})}{(T-t)S_t}}_{\frac{\mathbb{E}^*_t\left[(R_T - R_{f,t})\mathbf{1}\left\{R_T < R_{f,t}\right\}\right]}{T-t}} + R_{f,t}\,\text{down-SVIX}^2_t. \tag{29}$$

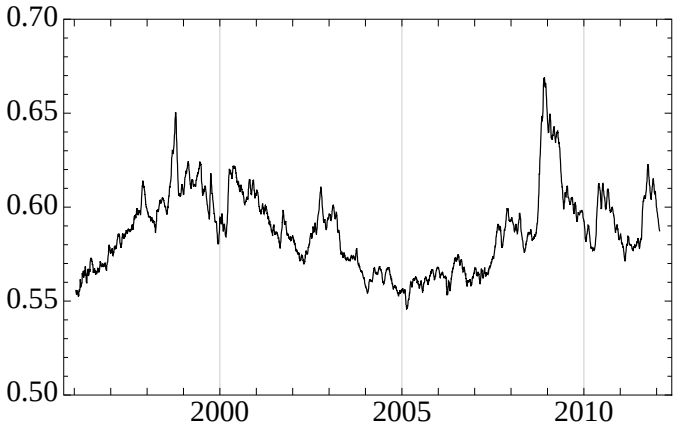
The *up-premium* $\frac{1}{T-t}\widetilde{\mathbb{E}}_t\left[R_T - R_{f,t} \mid R_T > R_{f,t}\right]$ is given by the ratio of (28) and (27); the *down-premium* $\frac{1}{T-t}\widetilde{\mathbb{E}}_t\left[R_T - R_{f,t} \mid R_T < R_{f,t}\right]$ can be calculated similarly from (29) and (27).

Figure 7 plots the log investor's perceived probability of an up-move at horizons $T = 1$ month and $T = 1$ year. At the 1-month horizon, the probability remains close to, though always above, 50%. The probability of an up-move is highest during the crisis in November 2008. (Shiller (1987) documents a similar phenomenon, that many investors expected a rebound following the crash of Black Monday.) At the 1-year horizon, the probabilities are further from (but still always above) 50%, as one would expect. Aside from late 2008, the late 1990s are revealed as a time when the subjective probability of an up-move was unusually high.

[18] As is well-known, the price of a digital call option that pays \$1 if $S_T > K$ and nothing otherwise is $-\mathrm{call}'_{t,T}(K)$; the proof is an exercise in the logic of static replication.



(a) 1 month

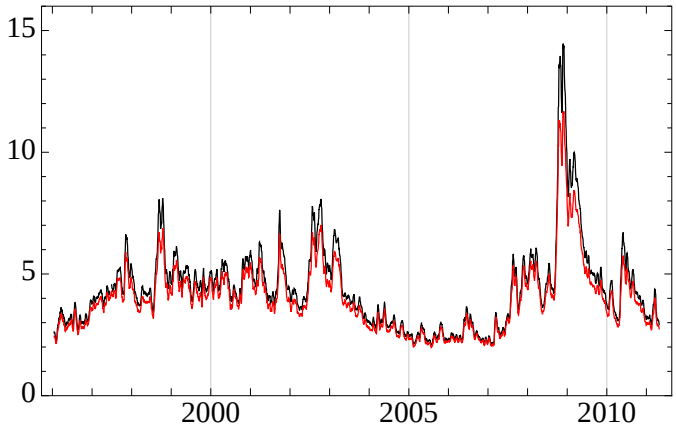


(b) 1 year

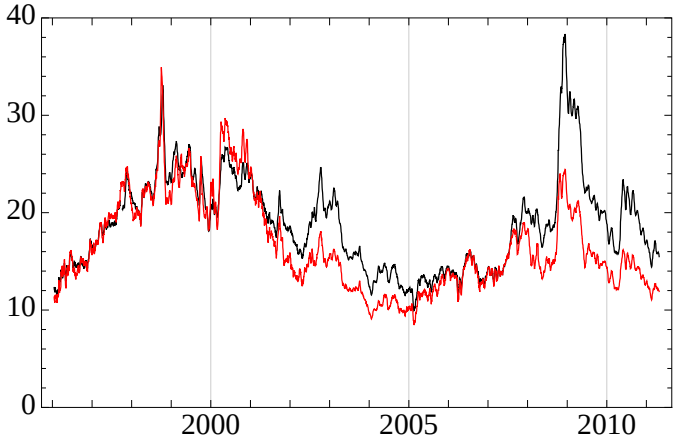
Figure 7: Log investor's perceived probability of an up-move, $\widetilde{\mathbb{P}}(R_T > R_{f,T})$.

Figure 8 plots the up-premium and down-premium over 1-month and 1-year horizons; the sign on the down-premium is flipped in both panels so that the up- and down-premium can be easily compared. At the 1-month horizon, the up-premium and down-premium hit peaks (in absolute value) in November 2008. Thus although I have argued that the equity premium was high at that time, and that the probability of an up-move was high at that time, the expected return on the market *conditional on a down-move* was more negative then than at any other point in my sample.

At the 1-year horizon, it is interesting to compare November 1998 with November 2008, as in Table 4. Both were times when the equity premium was high; when the probability of an up-move was high; and when the up-premium was high. The main



(a) 1 month



(b) 1 year

Figure 8: Black: The 'up-premium' $\widetilde{\mathbb{E}}\left(R_T - R_{f,T} \mid R_T > R_{f,T}\right)$. Red: the 'down-premium' $\widetilde{\mathbb{E}}\left(R_T - R_{f,T} \mid R_T < R_{f,T}\right)$ (sign flipped).

distinction between the two is that the down-premium was strikingly large in absolute value in November 1998. The late 1990s and early 2000s is the only period in my sample in which the equity premium was larger in magnitude conditional on a down-move than on an up-move, suggesting that sophisticated market participants were bullish, but appreciated that a major sell-off was possible (consistent with the findings of Brunnermeier and Nagel (2004) but using, of course, very different data).

	11/98	11/08
$\widetilde{\mathbb{E}}_t\left(R_T - R_{f,t}\right)$	11%	18%
$\widetilde{\mathbb{P}}_t\left(\text{up-move}\right)$	65%	67%
$\widetilde{\mathbb{E}}_t\left(R_T - R_{f,t} \mid \text{up-move}\right)$	33%	38%
$\widetilde{\mathbb{E}}_t\left(R_T - R_{f,t} \mid \text{down-move}\right)$	−35%	−24%

Table 4: A comparison of November 1998 and November 2008. T = one-year horizon.

8 Conclusion

The starting point of this paper is the identity (4), which shows that the expected excess return on any asset equals the risk-neutral variance of the asset's return minus a covariance term. If options are traded on the asset, then risk-neutral variance can be unambiguously measured without requiring any assumptions other than the absence of arbitrage. I apply the identity to the return on the market. In this case, risk-neutral variance is equal to the square of a volatility index, SVIX, that is similar to the VIX index, and I argue that the covariance term is weakly negative. The square of the SVIX index is therefore a lower bound on the equity premium.

I construct the SVIX index using S&P 500 index option data from 1996 to 2012. The index is strikingly volatile; it implies that in late 2008, the equity premium rose above 20% at the 1-year horizon and above 55% (annualized) at the 1-month horizon. More aggressively, I argue that the lower bound is approximately *tight*—that is, the SVIX index is not merely a lower bound on the equity premium, it is approximately *equal* to the equity premium.

These results point to a novel view of the equity premium. First, they suggest that the equity premium is far more volatile than implied by the valuation-ratio predictors

of Campbell and Thompson (2008). The distinction between the two views is sharpest on days such as Black Monday, in 1987, when the S&P 500 and Dow Jones indices experienced very severe declines, with daily returns roughly twice as negative as the next-worst day in history. On the Campbell–Thompson view of the world, the equity premium rose on the order of two or three percentage points during this episode. In contrast, option prices are known to have exploded on Black Monday, which I argue implies also that the equity premium exploded. Second, this volatility often reflects movements in the equity premium at weekly, daily, or even higher frequency. The macro-finance literature, which seeks to rationalize market gyrations at the business cycle frequency, typically has not acknowledged or attempted to address such movements. Third, the equity premium is strongly right-skewed: the median equity premium is on the order of 3 or 4%, but there are occasional opportunities for unconstrained investors to earn a much higher equity premium. Fourth, the term structure of the equity premium reveals that during these occasional episodes in which the equity premium is very high, a disproportionate fraction of the equity premium is concentrated in the form of high expected returns over the very short run.

9 References

Ang, A., and G. Bekaert (2007), "Stock Return Predictability: Is It There?" *Review of Financial Studies*, 20:3:651–707.

Bansal, R., D. Kiku, I. Shaliastovich, and A. Yaron (2012), "Volatility, the Macroeconomy, and Asset Prices," working paper.

Bansal, R. and A. Yaron (2004), "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles," *Journal of Finance*, 59:4:1481–1509.

Barro, R. J. (2006), "Rare Disasters and Asset Markets in the Twentieth Century," *Quarterly Journal of Economics*, 121:3:823–866.

Bates, D. S. (2008), "The Market for Crash Risk," *Journal of Economic Dynamics and Control*, 32:7:2291–2321.

Bates, D. S. (2012), "US Stock Market Crash Risk, 1926–2010," *Journal of Financial Economics*, forthcoming.

Ben-David, I., J. R. Graham, and C. R. Harvey (2013), "Managerial Miscalibration," *Quarterly Journal of Economics*, 128:1547–1584.

Black, F., and M. Scholes (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81:637–659.

Bollersley, T., G. Tauchen, and H. Zhou (2009), "Expected Stock Returns and Variance Risk Premia," *Review of Financial Studies*, 22:11:4463–4492.

Breeden, D. T., and R. H. Litzenberger (1978), "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, 51:4:621–651.

Brunnermeier, M. and S. Nagel (2004), "Hedge Funds and the Technology Bubble," *Journal of Finance*, 59(5), 2013-2040.

Campbell, J. Y. and J. H. Cochrane (1999), "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy*, 107:2:205–251.

Campbell, J. Y., S. Giglio, C. Polk, and R. Turley, "An Intertemporal CAPM with Stochastic Volatility," working paper.

Campbell, J. Y., and R. J. Shiller (1988), "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1:3:195–228.

Campbell, J. Y., and S. B. Thompson (2008), "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies*, 21:4:1509–1531.

Cochrane, J. H. (2005), *Asset Pricing*, Princeton University Press, Princeton, NJ.

Cochrane, J. H. (2008), "The Dog That Did Not Bark: A Defense of Return Predictability," *Review of Financial Studies*, 21:4:1533–1575.

Cochrane, J. H. (2011), "Discount Rates," *Journal of Finance*, 66:4:1047–1108.

Drechsler, I., and A. Yaron (2011), "What's Vol Got to Do with It," *Review of Financial Studies*, 24:1:1–45.

Epstein, L., and S. Zin (1989), "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework," *Econometrica*, 57:937–969.

Fama, E. F., and K. R. French (1988), "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, 22:3–25.

Fama, E. F., and K. R. French (1996), "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance*, 51:1:55–84.

Fama, E. F., and K. R. French (2002), "The Equity Premium," *Journal of Finance*, 57:2:637–659.

Goyal, A., and I. Welch (2008), "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21:4:1455–1508.

Greenwood, R., and A. Shleifer (2014), "Expectations of Returns and Expected Returns," *Review of Financial Studies*, 27:3:714–746.

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50:4:1029–1054.

Hansen, L. P. and R. Jagannathan (1991), "Implications of Security Market Data for Models of Dynamic Economies," *Journal of Political Economy*, 99:2:225–262.

Harvey, C. R., and A. Siddique (2000), "Conditional Skewness in Asset Pricing Tests," *Journal of Finance* 55:3:1263–1295.

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

Kraus, A., and R. H. Litzenberger (1976), "Skewness Preference and the Valuation of Risk Assets," *Journal of Finance* 31:4:1085–1100.

Lettau, M., and S. Ludvigson (2001), "Consumption, Aggregate Wealth, and Expected Stock Returns," *Journal of Finance*, 56:3:815–849.

Liu, J., J. Pan, and T. Wang (2005), "An Equilibrium Model of Rare-Event Premia and Its Implication for Option Smirks," *Review of Financial Studies*, 18:131–164.

Malmendier, U., and S. Nagel (2011), "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?" *Quarterly Journal of Economics*, 126:1:373–416.

Martin, I. W. R. (2013), "Simple Variance Swaps," working paper.

Merton, R. C. (1980), "On Estimating the Expected Return on the Market," *Journal of Financial Economics*, 8:323–361.

Pan, J. (2002), "The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study," *Journal of Financial Economics*, 63:3–50.

Roll, R. (1977), "A Critique of the Asset Pricing Theory's Tests I: On Past and Potential Testability of the Theory," *Journal of Financial Economics*, 4:129–176.

Ross, S. A. (2005), *Neoclassical Finance*, Princeton University Press.

Rubinstein, M. E. (1973), "The Fundamental Theorem of Parameter-Preference Security Valuation," *Journal of Financial and Quantitative Analysis*, 8:1:61–69.

Shiller, R. J. (1987), "Investor Behavior in the October 1987 Stock Market Crash: Survey Evidence," NBER Working Paper 2446.

Shiller, R. J. (2000), *Irrational Exuberance*, Princeton University Press.

Stambaugh, R. F. (1999), "Predictive Regressions," *Journal of Financial Economics*, 54:375–421.

Wachter, J. (2013), "Can time-varying risk of rare disasters explain aggregate stock market volatility?" *Journal of Finance*, 68:987–1035

A The negative correlation condition

This section contains proofs that the examples in Section 2 satisfy the negative correlation condition (NCC).

Example 1. Write $M_T = e^{-r_{f,t}+\sigma_{M,t}Z_{M,T}-\sigma_{M,t}^2/2}$ and $R_T = e^{\mu_{R,t}+\sigma_{R,t}Z_{R,T}-\sigma_{R,t}^2/2}$, where $Z_{M,T}$ and $Z_{R,T}$ are (potentially correlated) standard Normal random variables. Define $\lambda_t = (\mu_{R,t} - r_{f,t})/\sigma_{R,t}$ to be the Sharpe ratio conditional on time-t information. Some straightforward algebra shows that $\mathbb{E}_t\, M_T R_T^2 \leq \mathbb{E}_t\, R_T$ if and only if $\lambda_t \geq \sigma_{R,t}$.

Example 2. By assumption, there is an investor with wealth W_t and utility function $u(\cdot)$ who chooses, at time t, from the available menu of assets with returns $R_T^{(i)}$, $i = 1, 2, \ldots$. In other words, he chooses portfolio weights $\{w_i\}$ to solve the problem

$$\max_{\{w_i\}} \mathbb{E}_t\, u\left[W_t\left(\sum_i w_i R_T^{(i)}\right)\right] \quad \text{subject to} \quad \sum_i w_i = 1. \tag{30}$$

The first-order condition for (say) w_j is that

$$\mathbb{E}_t\left[W_t u'\left(W_t \sum_i w_i R_T^{(i)}\right) R_T^{(j)}\right] = \lambda_t,$$

where $\lambda_t > 0$ is the Lagrange multiplier associated with the constraint in (30). Since the investor chooses to hold the market, we have $\sum_i w_i R_T^{(i)} = R_T$. Thus,

$$\mathbb{E}_t \left[\underbrace{\frac{W_t}{\lambda_t} u'\left(W_t R_T\right)}_{M_T} R_T^{(j)} \right] = 1$$

for any return $R_T^{(j)}$. It follows that the SDF is proportional (with a constant of proportionality that is known at time t) to $u'(W_t R_T)$.

To show that the NCC holds, we must show that $\mathrm{cov}_t(u'(W_t R_T) R_T, R_T) \leq 0$. This holds because $u'(W_t R_T) R_T$ is decreasing in R_T: its derivative is $u'(W_t R_T) + W_t R_T u''(W_t R_T) = -u'(W_t R_T)\left[\gamma(W_t R_T) - 1\right]$, which is negative because risk aversion $\gamma(x) \equiv -x u''(x)/u'(x)$ is at least one.

If the investor has log utility, then $\gamma(x) \equiv 1$, so the inequality holds with equality. But it is not *necessary* for the investor to have log utility for the inequality to hold with equality: all we require is that $M_T R_T$ is uncorrelated with R_T. That is, we merely need that $M_T = I_T / R_T$ where I_T and R_T are uncorrelated (and $\mathbb{E}_t I_T = 1$ since $\mathbb{E}_t M_T R_T$ must equal one). Log utility is the special case in which $I_T \equiv 1$.

Examples 3a and 3b. For reasons given in the text, Example 3a is a special case of Example 3b, which we now prove. We must check that $\mathrm{cov}_t(M_T R_T, R_T) \leq 0$, or equivalently that

$$\mathrm{cov}_t(-R_T V_W(W_T, z_{1,T}, \ldots, z_{N,T}), R_T) \geq 0. \tag{31}$$

That is, we must prove that the covariance of two functions of $R_T, R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}$ is positive. The two functions are

$$f(R_T, R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}) = -R_T V_W(\alpha_t(W_t - C_t) R_T + (1-\alpha_t)(W_t - C_t) R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}) \tag{32}$$

and

$$g(R_T, R_T^{(i)}, z_{1,T}, \ldots, z_{N,T}) = R_T.$$

(Since the covariance is conditional on time-t information, α_t and $(W_t - C_t)$ can be treated as known constants.) By the defining property of associated random variables, (31) holds so long as f and g are each weakly increasing functions of their arguments. This is obviously true for g, so it only remains to check that the first derivatives of f are all nonnegative.

Differentiating (32) with respect to R_T, we need $-V_W(W_T, z_{1,T}, \ldots, z_{N,T}) - \alpha_t(W_t - C_t)R_T V_{WW}(W_T, z_{1,T}, \ldots, z_{N,T}) \geq 0$, or equivalently

$$-\frac{W_T V_{WW}(W_T, z_{1,T}, \ldots, z_{N,T})}{V_W(W_T, z_{1,T}, \ldots, z_{N,T})} \geq \frac{W_T}{W_{M,T}},$$

where W_T and $W_{M,T}$ are as given in the main text. This is the constraint on risk aversion.

Differentiating (32) with respect to $R_T^{(i)}$, we need $-R_T(1-\alpha_t)(W_t - C_t)V_{WW}(W_T, z_{1,T}, \ldots, z_{N,T}) \geq 0$, which follows because $V_{WW} < 0$.

Differentiating (32) with respect to $z_{j,T}$, we need $-R_T V_{Wj}(W_T, z_{1,T}, \ldots, z_{N,T}) \geq 0$, which follows because V_{Wj} (the cross derivative of the value function with respect to wealth and the jth state variable) is weakly negative due to the choice of sign on the state variables.

Examples 4a and 4b. With Epstein–Zin preferences, the SDF is proportional (up to quantities known at time t) to $(W_T/C_T)^{(\gamma-1)/(1-\psi)} R_T^{-\gamma}$, so the desired inequality, $\mathrm{cov}_t(M_T R_T, R_T) \leq 0$, is equivalent to

$$\mathrm{cov}_t \left[-\left(\frac{W_T}{C_T}\right)^{(\gamma-1)/(1-\psi)} R_T^{1-\gamma}, R_T \right] \geq 0.$$

If $\gamma = 1$, as in Example 4b, then this holds with equality.

If W_T/C_T and R_T are associated, as assumed in Example 4a, then we need to check the first derivatives of

$$f(x, y) = -x^{(\gamma-1)/(1-\psi)} y^{1-\gamma}$$

to be nonnegative. That is, we need $\gamma \geq 1$ and $\psi \geq 1$, as claimed.

B Calculating risk-neutral variance

Note that for any x, we have

$$x^2 = 2 \int_0^\infty \max\{0, x - K\}\, dK.$$

Setting $x = S_T$, taking risk-neutral expectations, and multiplying by $\frac{1}{R_{f,t}}$,

$$\begin{aligned} \frac{1}{R_{f,t}} \mathbb{E}_t^* S_T^2 &= 2 \int_0^\infty \frac{1}{R_{f,t}} \mathbb{E}_t^* \max\{0, S_T - K\}\, dK \\ &= 2 \int_0^\infty \mathrm{call}_{t,T}(K)\, dK. \end{aligned} \tag{33}$$

Using equations (9), (10), and (33), the risk-neutral variance can be calculated from observable prices:

$$\frac{1}{R_{f,t}} \operatorname{var}_t^* R_T = \frac{1}{S_t^2} \left[2 \int_0^\infty \operatorname{call}_{t,T}(K)\, dK - \frac{F_{t,T}^2}{R_{f,t}} \right]. \tag{34}$$

This expression incorporates the prices of in-the-money calls, which are usually illiquid. But by put-call parity, $\operatorname{call}_{t,T}(K) = \operatorname{put}_{t,T}(K) + \frac{1}{R_{f,t}}(F_{t,T} - K)$, so

$$\begin{aligned}
\int_0^\infty \operatorname{call}_{t,T}(K)\, dK &= \int_0^{F_{t,T}} \operatorname{call}_{t,T}(K)\, dK + \int_{F_{t,T}}^\infty \operatorname{call}_{t,T}(K)\, dK \\
&= \int_0^{F_{t,T}} \operatorname{put}_{t,T}(K) + \frac{1}{R_{f,t}}(F_{t,T} - K)\, dK + \int_{F_{t,T}}^\infty \operatorname{call}_{t,T}(K)\, dK \\
&= \int_0^{F_{t,T}} \operatorname{put}_{t,T}(K)\, dK + \frac{F_{t,T}^2}{2R_{f,t}} + \int_{F_{t,T}}^\infty \operatorname{call}_{t,T}(K)\, dK.
\end{aligned}$$

Substituting this into (34), we have the formula (13) for risk-neutral variance:

$$\frac{1}{R_{f,t}} \operatorname{var}_t^* R_T = \frac{2}{S_t^2} \left[\int_0^{F_{t,T}} \operatorname{put}_{t,T}(K)\, dK + \int_{F_{t,T}}^\infty \operatorname{call}_{t,T}(K)\, dK \right].$$

B.1 Construction of the lower bound

The data are from *OptionMetrics*, runnning from January 4, 1996, to January 31, 2012; they include the closing price of the S&P 500 index, and the expiration date, strike price, highest closing bid and lowest closing ask of all call and put options with fewer than 550 days to expiry. I clean the data in several ways. First, I delete all replicated entries (of which there are more than 500,000). Second, for each strike, I select the option—call or put—whose mid price is lower. Third, I delete all options with a highest closing bid of zero. Finally, I delete all Quarterly options, which tend to be less liquid than regular S&P 500 index options and to have a smaller range of strikes. Having done so, I am left with 1,165,585 option-day datapoints. I compute mid-market option prices by averaging the highest closing bid and lowest closing ask, and using the resulting prices to compute the lower bound by discretizing the right-hand side of inequality (16).

On any given day, I compute the lower bound at a range of time horizons depending on the particular expiration dates of options traded on that day, with the constraint that the shortest time to expiry is never allowed to be less than 7 days; this is the

same procedure that the CBOE follows. I then calculate the implied bound for $T =$ 30, 60, 90, 180, and 360 days by linear interpolation. Occasionally, extrapolation is necessary, for example when the nearest-term option's time-to-maturity first dips below 7 days, requiring me to use the two expiry dates further out; again, this is the procedure followed by the CBOE.

B.2 The effect of discrete strikes

The integrals that appear throughout the paper are idealizations: in practice we only observe options at some finite set of strikes. Write $\Omega_{t,T}(K)$ for the price of an out-of-the-money option with strike K, that is,

$$\Omega_{t,T}(K) \equiv \begin{cases} \text{put}_{t,T}(K) & \text{if } K < F_{t,T} \\ \\ \text{call}_{t,T}(K) & \text{if } K \geq F_{t,T} \end{cases};$$

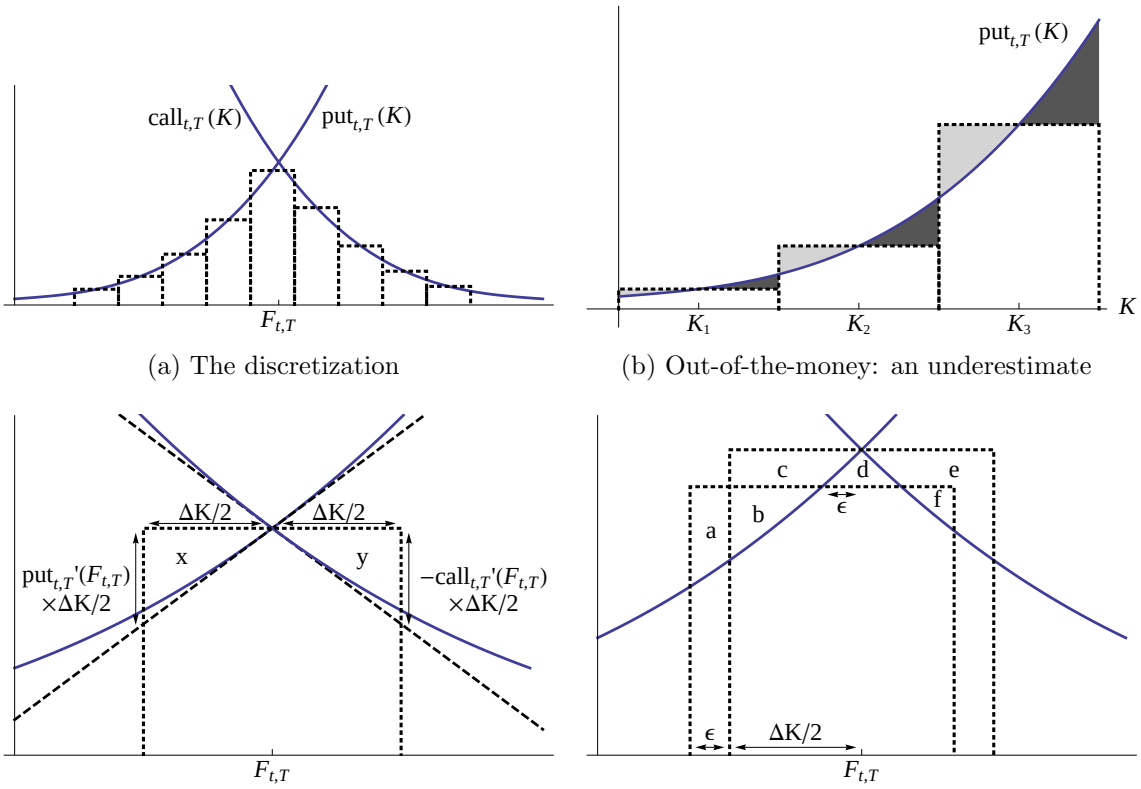
write $K_1, \ldots, K_N$ for the strikes of observable options; write K_j for the strike that is nearest to the forward price $F_{t,T}$;[19] and define $\Delta K_i \equiv (K_{i+1} - K_{i-1})/2$. Then the idealized integral $\int_0^\infty \Omega_{t,T}(K)\, dK$ is replaced, in practice, by the observable sum $\sum_{i=1}^N \Omega_{t,T}(K_i)\, \Delta K_i$. (This is the CBOE's procedure in calculating VIX, and I follow it in this paper.) Figure 9a illustrates.

The question is, how well does the sum approximate the integral? The next result shows that there are two forces pushing in the direction of underestimation (of the integral by the sum) and one pushing in the direction of overestimation. But the latter effect is very minor in practice, so one should think of discretization as leading to underestimation of the integral.

Result 1 (The effect of discretization by strike)**.** *Discretizing by strike will tend to lead to an underestimate of the idealized lower bound, in that*

$$\underbrace{\frac{2}{(T-t)R_{f,t}S_t^2}\sum_{i=1}^{N}\Omega_{t,T}(K_i)\,\Delta K_i}_{\textit{discretization}} \leq \underbrace{\frac{2}{(T-t)R_{f,t}S_t^2}\int_0^\infty \Omega_{t,T}(K)\,dK}_{\textit{idealized lower bound}} + \underbrace{\frac{(\Delta K_j)^2}{4(T-t)\cdot R_{f,t}^2\cdot S_t^2}}_{\textit{very small}}.$$

[19] For simplicity, I assume that strikes are evenly spaced near-the-money, $K_{j+1} - K_j = K_j - K_{j-1}$. This is not essential, but it is almost always the case in practice and lets me economize slightly on notation.



(a) The discretization

(b) Out-of-the-money: an underestimate

(c) At-the-money: an overestimate

(d) The worst-case scenario in Lemma 1

Figure 9: The effect of discretization. Different panels use different scales.

Proof. Non-observability of deep-out-of-the-money options obviously leads to an underestimate of the lower bound.

Consider, first, the out-of-the-money puts with strikes $K_1, \ldots, K_{j-1}$. The situation is illustrated in Figure 9b: by convexity of $\text{put}_{t,T}(K)$, the light grey areas that are included (when they should be excluded) are smaller than the dark grey areas that are excluded (when they should be included). The same logic applies to the out-of-the-money calls with strikes $K_{j+1}, K_{j+2}, \ldots$. Thus the observable options—excluding the nearest-the-money option—will always underestimate the part of the integral which they are intended to approximate.

It remains to consider the nearest-the-money option with strike K_j, which alone can lead to an overestimate. Lemma 1, below, shows that the worst case is if the strike of the nearest-the-money option happens to be exactly *equal* to the forward price $F_{t,T}$, as in Figure 9c. For an upper bound on the overestimate in this case we must find an upper bound on the sum of the approximately triangular areas (x) and (y) that are shown in the figure. We can do so by replacing the curved lines in the figure by the

(dashed) tangents to $\text{put}_{t,T}(K)$ and $\text{call}_{t,T}(K)$ at $K = F_{t,T}$. The areas of the resulting triangles provide the desired upper bound, by convexity of $\text{put}_{t,T}(K)$ and $\text{call}_{t,T}(K)$: we have

$$\text{area (x)} + \text{area (y)} \leq \frac{1}{2}\left(\frac{\Delta K}{2}\right)^2 \text{put}'_{t,T}(K) - \frac{1}{2}\left(\frac{\Delta K}{2}\right)^2 \text{call}'_{t,T}(K).$$

But, by put-call parity, $\text{put}'_{t,T}(K) - \text{call}'_{t,T}(K) = 1/R_{f,t}$. Thus, the overestimate due to the at-the-money option is at most

$$\frac{1}{2}\left(\frac{\Delta K}{2}\right)^2 \frac{1}{R_{f,t}}.$$

Since the contributions from out-of-the-money and missing options led to underestimates, the overall overestimate is at most this amount. Finally, since the definition scales the integral by $2/((T-t)R_{f,t}S_t^2)$, the result follows. □

The maximal overestimate provided by this result is *extremely small*: for the S&P 500 index, the interval between strikes near-the-money is $\Delta K_j = 5$. If, say, the forward price of the S&P 500 index is $F_{t,T} = 1000$ and we are considering a monthly horizon, $T - t = 1/12$, then the discretization leads to an overestimate of SVIX^2 that is *at most* $7.5 \times 10^{-5} < 0.0001$. By comparison, the average level of SVIX^2 is on the order of 0.05, as shown in Table 2. Since the non-observability of deep-out-of-the-money options causes underestimation, there is therefore a very strong presumption that the sum underestimates the integral.

It only remains to establish the following lemma, which is used in the proof of Result 1. The goal is to consider the largest possible overestimate that the option whose strike is nearest to the forward price, $F_{t,T}$, can contribute. Figure 9d illustrates. The dotted rectangle in the figure is the contribution if the strike happens to be *equal* to $F_{t,T}$; I will call this *Case 1*. The dashed rectangle is the contribution if the strike equals $F_{t,T} - \varepsilon$, for some $\varepsilon > 0$ (for concreteness—the case $\varepsilon < 0$ is essentially identical); I will call this *Case 2*.

Lemma 1. *The option with strike closest to the forward overestimates most in the case in which its strike is equal to the forward.*

Proof. The overestimate in Case 1 is greater than that in Case 2 if

$$\text{area (b)} + \text{area (c)} + \text{area (e)} + \text{area (f)} \geq \text{area (a)} + \text{area (b)} + \text{area (f)} - \text{area (d)}$$

in Figure 9d, or equivalently,

$$\text{area (c)} + \text{area (d)} + \text{area (e)} \geq \text{area (a)}. \tag{35}$$

But, by convexity of $\text{put}_{t,T}(K)$,

$$\text{area (b)} + \text{area (c)} \geq \text{area (a)} + \text{area (b)},$$

from which (35) follows. An almost identical argument applies if $\varepsilon < 0$. □

C Tables

Table 5 reports GMM results corresponding to those in Table 1, using the inverse of the estimated covariance matrix of pricing errors as the weighting matrix in a second stage. This was shown to be asymptotically optimal by Hansen (1982), but Cochrane (2005) argues that it may be less robust than the approach adopted in the main text in short finite samples. In the present case, the results are very similar with either approach.

Table 6 reproduces the results in Table 3, but excludes the period August 1, 2008–July 31, 2009.

Table 7 reports results for regressions

$$R_T - R_{f,t} = \alpha + \beta_1 \times R_{f,t} \cdot \text{SVIX}_t^2 + \beta_2 \times VRP_t + \varepsilon_T \tag{36}$$

of realized returns onto risk-neutral variance and a measure of the variance risk premium, $VRP_t \equiv R_{f,t} \cdot \text{SVIX}_t^2 - \text{SVAR}_t$. Realized daily return variance, SVAR_t, is computed at time t by looking backwards over the same horizon-length, $T - t$, as the corresponding forward-looking realized return (so, for example, I use 1-month backward-looking realized variances to predict 1-month forward-looking realized returns). If realized variance is a good proxy for forward-looking real-world variance, this is a measure of the 'variance risk premium.'

Consistent with the empirical findings of Bollerslev, Tauchen and Zhou (2009) and Drechsler and Yaron (2011), the coefficient on VRP_t is positive and strongly significant at predictive horizons out to 3 months.[20] This predictive success reflects the fact that

[20]My approach follows that of Bollerslev, Tauchen and Zhou (2009) rather than that of Drechsler and Yaron (2011), who use predictive regressions to forecast the evolution of variance itself. I follow the former approach to avoid in-sample/out-of-sample issues.

	constant	$R_M - R_f$	SMB	HML	MOM	$\widehat{\text{cov}}(M_T R_T, R_T)$
Full sample	1.013	−1.235	−1.463	−2.068	—	−0.0027
	(0.007)	(0.618)	(0.870)	(0.810)	—	(0.0016)
Jul '26–Dec '62	1.013	−1.355	−0.896	−0.420	—	−0.0032
	(0.010)	(0.925)	(1.215)	(1.326)	—	(0.0028)
Jan '63–Feb '14	1.055	−3.475	−2.962	−8.018	—	−0.0030
	(0.019)	(1.120)	(1.426)	(1.574)	—	(0.0020)
Jan '96–Feb '14	1.037	−2.524	−3.208	−5.726	—	−0.0024
	(0.028)	(1.709)	(2.154)	(2.298)	—	(0.0034)
Full sample	1.078	−2.778	−1.592	−5.119	−5.584	−0.0033
	(0.019)	(0.701)	(0.989)	(1.052)	(0.957)	(0.0019)
Jan '27–Dec '62	1.096	−2.986	−1.340	−3.968	−6.961	−0.0027
	(0.030)	(0.976)	(1.346)	(1.775)	(1.466)	(0.0036)
Jan '63–Dec '13	1.110	−4.576	−4.343	−9.920	−4.868	−0.0038
	(0.029)	(1.268)	(1.371)	(1.710)	(1.254)	(0.0022)
Jan '96–Dec '13	1.076	−4.387	−3.230	−7.976	−3.349	−0.0031
	(0.039)	(1.987)	(2.110)	(2.356)	(1.533)	(0.0036)

Table 5: Estimates of coefficients in the factor models (7) (upper panel) and (8) (lower panel), and of $\text{cov}(M_T R_T, R_T)$, using the inverse of the estimated covariance matrix of pricing errors as the weighting matrix.

horizon	$\widehat{\alpha}$	s.e.	$\widehat{\beta}$	s.e.	R^2
1 mo	−0.095	[0.061]	3.705	[1.258]	3.36%
2 mo	−0.081	[0.062]	3.279	[1.181]	4.83%
3 mo	−0.076	[0.067]	3.147	[1.258]	5.98%
6 mo	−0.043	[0.072]	2.319	[1.276]	4.94%
1 yr	0.045	[0.088]	0.473	[1.731]	0.27%

Table 6: Coefficient estimates for the regression (18), excluding the crisis period August 1, 2008–July 31, 2009 from the sample.

horizon	$\widehat{\alpha}$	s.e.	$\widehat{\beta}_1$	s.e.	$\widehat{\beta}_2$	s.e.	R^2
1 mo	−0.086	[0.063]	2.048	[1.273]	3.908	[1.053]	4.96%
2 mo	−0.113	[0.061]	2.634	[1.007]	3.884	[0.761]	8.54%
3 mo	−0.086	[0.071]	2.273	[1.407]	2.749	[0.346]	6.79%
6 mo	−0.051	[0.076]	1.992	[1.132]	−0.525	[1.259]	6.56%
1 yr	−0.073	[0.078]	2.278	[0.909]	−0.694	[0.680]	10.34%

Table 7: Coefficient estimates for the regression (36).

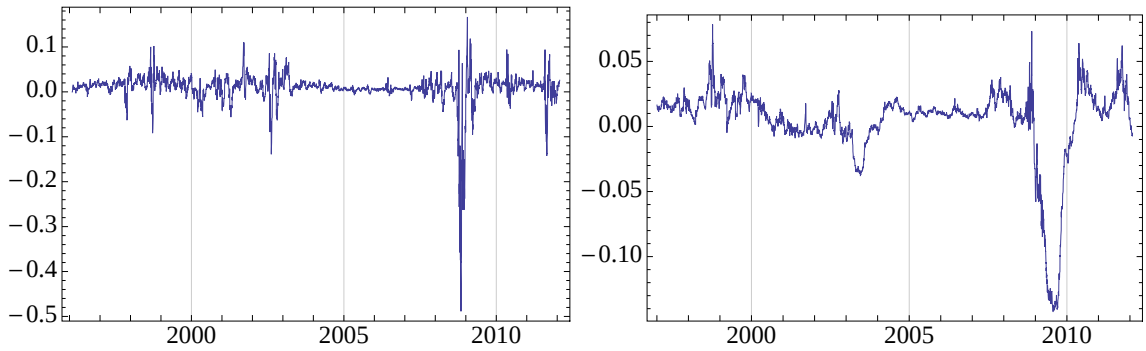
horizon	$\widehat{\alpha}$	s.e.	$\widehat{\beta}_1$	s.e.	$\widehat{\beta}_2$	s.e.	R^2
1 mo	−0.103	[0.061]	3.333	[1.292]	1.548	[1.125]	3.61%
2 mo	−0.097	[0.063]	3.137	[1.353]	1.532	[1.801]	6.04%
3 mo	−0.083	[0.068]	2.902	[1.451]	1.133	[1.855]	6.34%
6 mo	0.016	[0.071]	0.797	[1.560]	0.360	[2.095]	0.74%
1 yr	0.008	[0.061]	0.331	[2.274]	1.761	[3.760]	3.10%

Table 8: Coefficient estimates for the regression (36), excluding the crisis period August 1, 2008 to July 31, 2009.

implied and realized volatility, SVIX_t and SVAR_t, rose sharply as the S&P 500 dropped in late 2008; implied volatility then fell relatively quickly, while SVAR_t declined more sluggishly. VRP_t therefore turned dramatically negative in late 2008, as shown in Figure 10 below. Since the market then continued to fall, this sluggish response of VRP_t helps fit the data. At the 6-month and 1-year horizons, however, VRP_t responds too sluggishly—it remains strongly negative even as the market starts to rally in March, 2009—so there is a sign-flip, with *negative* estimates of the coefficient on VRP_t are *negative* at the 6-month and 1-year horizons. The empirical facts are therefore hard to interpret: the sign-flip raises the concern that the apparent success of VRP_t as a predictor variable may be an artefact of this particular sample period. Table 8 therefore repeats the regression (36), but excludes the period from August 1, 2008 to July 31, 2009. Once this crisis period is excluded, VRP_t does not enter significantly at any horizon.

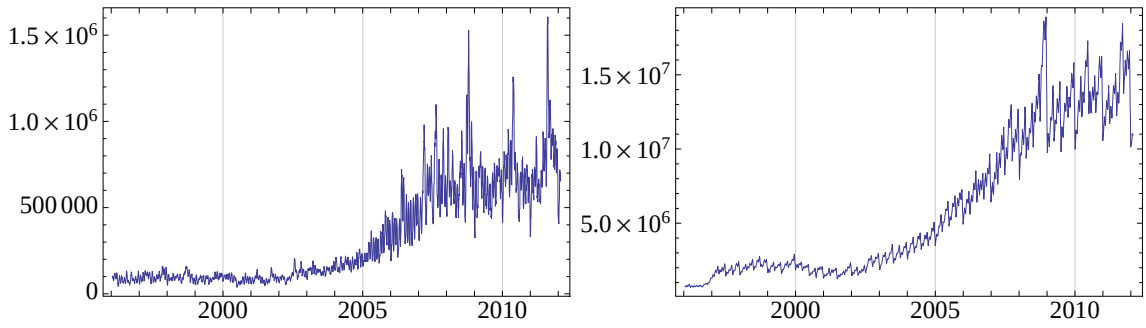
From a theoretical point of view, it is hard to rationalize a negative equity pre-

mium forecast within any equilibrium model. It is also implausible that the correctly-measured variance risk premium should ever be negative. More specifically, Bollerslev, Tauchen and Zhou (2009) show that within their own preferred equilibrium model, the variance risk premium would always be positive.



(a) 1-month (b) 1-year

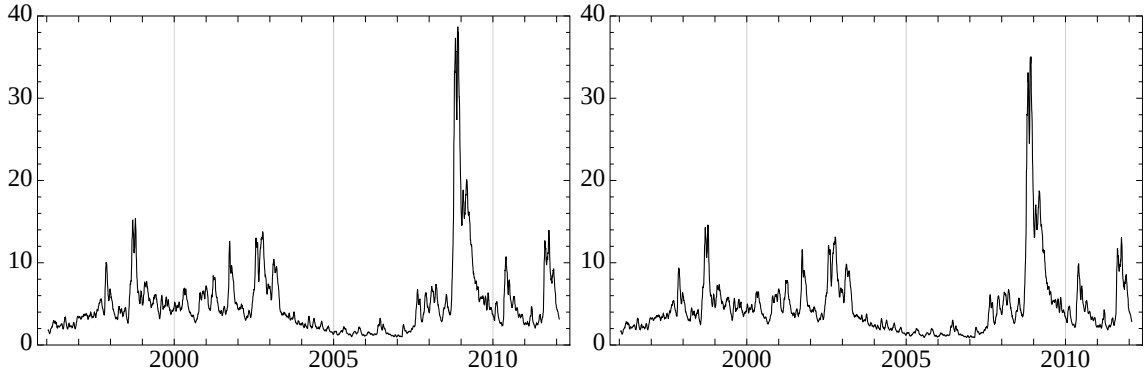
Figure 10: The variance risk premium, calculated as $R_{f,t} \cdot \text{SVIX}_t^2 - \text{SVAR}_t$.



(a) Volume (b) Open interest

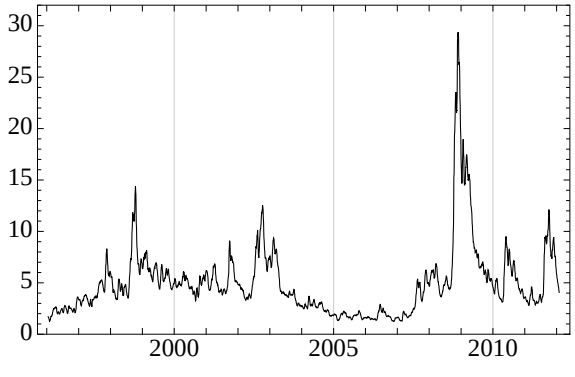
Figure 11: Volume and open interest in S&P 500 index options. The figures show 10-day moving averages.

The calculations of R_{OS}^2 in Table 3 depend on the rolling mean historical equity premium shown in Figure 14. The rolling mean is computed using the data series used by Campbell and Thompson (2008), which itself is based on S&P 500 total returns from February 1871, with the data prior to January 1927 obtained from Robert Shiller's website.

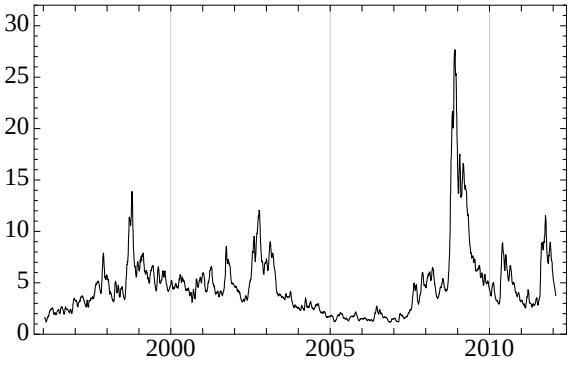


(a) 1 month, mid prices

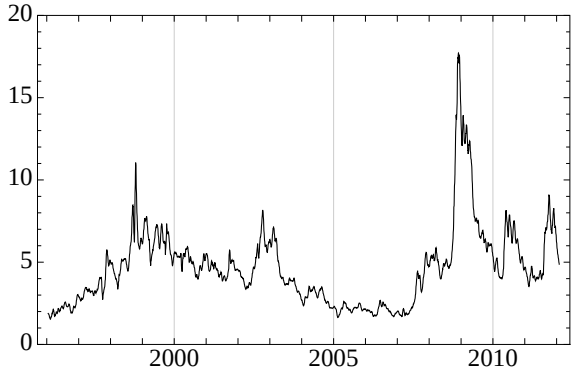
(b) 1 month, bid prices



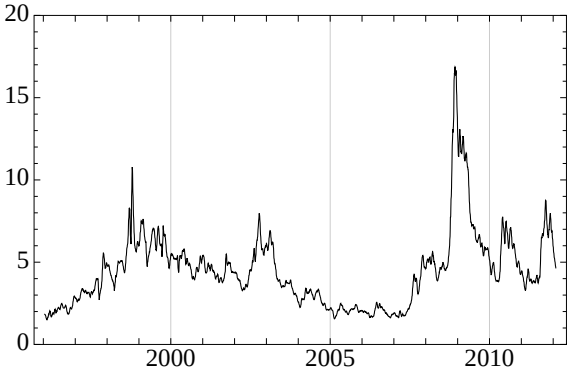
(c) 3 month, mid prices



(d) 3 month, bid prices



(e) 1 year, mid prices



(f) 1 year, bid prices

Figure 12: The lower bound on the annualized equity premium at different horizons (in %). The figures show 10-day moving averages. Mid prices on left; bid prices on right.

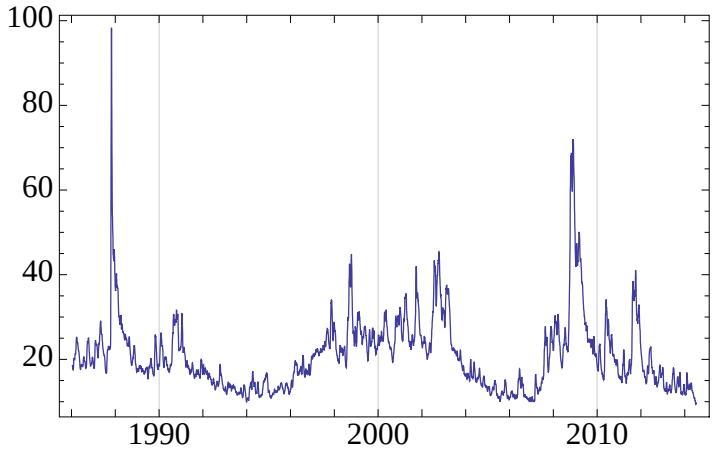


Figure 13: The VXO index, which exploded on Black Monday, October 19, 1987. 10-day moving average.

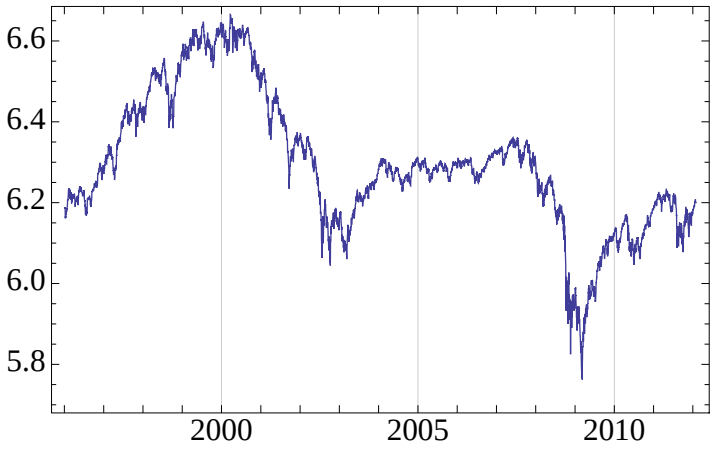


Figure 14: Historical mean equity premium, annualized, on a rolling basis.