Volatility Expectations and Returns

August 30, 2019

Abstract

We provide evidence that agents have slow moving beliefs about stock market volatility. This is supported in survey data and is also reflected in firm level option prices. We embed these expectations into a general equilibrium asset pricing model and show that we jointly explain the following stylized facts (some of which are novel to this paper): when volatility increases the equity premium and variance risk premium fall or stay flat at short horizons, despite higher future risk; these premiums appear to rise at longer horizons after future volatility has subsided; strategies that time volatility generate alpha; the variance risk premium forecasts stock returns more strongly than either realized variance or risk-neutral variance (VIX); changes in volatility are negatively correlated with contemporaneous returns. Slow moving expectations about volatility lead agents to initially underreact to volatility news followed by a delayed overreaction. This results in a weak, or even negative, risk-return tradeoff at shorter horizons but a stronger tradeoff at longer horizons (beyond where one can strongly forecast volatility). These dynamics are mirrored in the VIX and variance risk premium which reflect investor expectations about volatility.

1. Introduction

The link between risk and return is at the core of many asset pricing models, though there is weak empirical evidence of a risk-return tradeoff over time in the stock

*

market.[1] For example, consider the canonical mean-variance representative agent model equilibrium[2]

$$E_t[r_{t+1}] - r_{f,t} = \gamma\sigma_t^2 \tag{1}$$

where γ is the representative agents risk-aversion, σ_t^2 is the conditional variance of the market and $E_t[r_{t+1}] - r_{f,t}$ is the (log) market risk premium. This equation implies a tight link between volatility and expected returns. While this model may seem like a simplistic "straw man," the strong positive link between conditional volatility and expected returns is a fairly general feature of leading structural equilibrium asset pricing models in the literature.[3]

Empirically a long literature finds that the relationship between conditional variance and risk premia is weak at best. In particular, measures of conditional stock market variance or volatility do not strongly positively forecast returns so that risk-return ratios weaken when volatility rises (Glosten et al., 1993; Moreira and Muir, 2017, 2019). In fact, we show that the change in conditional volatility over the period of a few months if anything appears to negatively forecasts future stock returns over the next month, rather than strongly positively forecasting returns as in the basic model. However, we also show that longer term lags of volatility do appear to predict returns with a positive sign. Yet, these lags of volatility forecast returns at horizons for which they no longer strongly forecast future volatility, especially once controlling for more recent lags. Hence they no longer represent expected volatility – the object that should be linked to expected returns in theory. Measures of the variance risk premium display strikingly similar patterns. For example, increases in volatility if anything negatively predict the premium on VIX futures at short horizons of up to a month (Cheng, 2018) and we show this is true for variance swaps and other measures of the variance risk premium as well. That is, claims that provide insurance against future volatility, which are unconditionally expensive, appear "too cheap" after volatility rises. Similar to the pattern in stock returns, increases in volatility positively forecast the variance risk premium at longer horizons, peaking around six months.

While the relation between conditional volatility and expected returns is weak, there are some indirect facts potentially more in favor of a risk-return tradeoff. First,

[1] See, for example, Glosten, Jagannathan, and Runkle (1993) or Moreira and Muir (2017).

[2] E.g., Merton (1980)

[3] Martin (2016) argues this relationship is very general if σ_t^2 is replaced by risk-neutral variance which we will consider empirically as well. Moreira and Muir (2017) show that the basic risk-return relation is very strong in calibrations of leading equilibrium asset pricing models (including rational, behavioral, and those involving intermediation and frictions).

realized stock returns are highly negatively related to contemporaneous innovations in volatility (French, Schwert, and Stambaugh, 1987). This strong negative relation suggests a discount rate effect, since a higher discount rate in response to higher volatility pushes current stock prices – and hence realized returns – lower. However, it is of course puzzling if this effect is a discount rate effect that returns next period are not higher. Second, the variance risk premium (VIX minus realized variance) *does* strongly forecast stock returns suggesting equity risk premia are high when the variance risk premium is high (Bollerslev, Tauchen, and Zhou, 2009).[4]

The goal of this paper is to propose a model which jointly tackles these dynamics between realized volatility, the VIX, the variance risk premium, and stock returns. We do so by making one change to an otherwise standard Epstein Zin equilibrium model with stochastic volatility: we allow the representative agent to have slow moving expectations about volatility, which we show is supported by survey evidence. In our model agents form beliefs about volatility by taking a weighted average of past volatility observations. When the agent pays relatively too much attention to past, rather than current, volatility, they temporarily underreact to increases in volatility and then subsequently overreact. When agents see volatility increase they do react partially, driving prices down so that volatility is associated with negative returns contemporaneously. However, the initial underreaction means prices can continue to fall in the next period making ex-ante "risk premiums" appear weak, or even negative, and the subsequent overreaction keeps prices depressed for longer before eventually bouncing back, making it appear as though risk premiums are high well after the shock to volatility has largely subsided.[5] Market expectations of volatility (the VIX) mirror this, meaning the variance risk premium can initially fall before slowly rising at longer horizons. Thus the model matches the conditional dynamics of both equity and variance risk premiums following an increase in volatility, though we show this reconciles several additional pieces of evidence as well.

To see how the model delivers these dynamics, suppose that the true volatility process follows an AR(1) as in our model. This means the agents' objective best guess for next months volatility only depends on current volatility. However, suppose agents form a forecast for next months volatility based on an average of volatility over the past several months. Then, if volatility increases, agents expectations of

[4]See also Drechsler and Yaron (2011)

[5]We use the term risk premium as an empirical observation about the behavior of expected returns though in our model variance risk premiums only move because of agents biased expectations. Similarly, and in keeping with the literature, we often use "risk-nuetral variance" to refer to the VIX though much of the behavior in VIX in our model is driven by biased beliefs rather than standard risk-pricing.

volatility will increase somewhat but will underreact to the news – agents will not update their forecast about next periods volatility strongly enough since they average across several lags of volatility. However, they will still demand a higher equity premium due to the higher subjective expectation of volatility, hence stock prices will fall, generating a strong negative correlation between stock returns and volatility innovations (French et al., 1987). Similarly, "risk-neutral" expectations of volatility (the VIX) will rise, though also not strongly enough relative to a rational forecast. We use the term "risk-neutral" in keeping with terminology in the literature, though in our setting it reflects expectations under the agents potentially subjective beliefs. Because the VIX does not initially rise strongly enough, VIX minus the rational expectation of realized volatility (the variance risk premium) will fall, thus volatility news will negatively forecast the variance risk premium as is true empirically (see Cheng (2018), which we extend, as well as Poteshman (2001)).

When the next period arrives, suppose for simplicity there is no additional news about future volatility. With rational expectations, expected volatility declines as volatility mean reverts toward the unconditional average, and the equity premium would decline with expected volatility while stock prices rise on average. Instead, in our model, the agent may again update his expectation about volatility because he averages two periods of relatively elevated volatility, and may require a relatively higher equity premium. If the initial underreaction is strong enough, this pushes equity prices down further through an additional discount rate effect. Through this channel, the ex-ante news about higher volatility in the previous period can appear to forecast negative stock returns in the next period, hence the initial innovation in volatility can negatively forecast returns in the short term. Thus, the risk return tradeoff appears weak or even negative (Glosten et al., 1993), and this leads to apparently profitable volatility timing strategies (Moreira and Muir, 2017, 2019). This also nicely reconciles the puzzling evidence that shocks to volatility do line up with contemporaneous drops in stock prices, consistent with a discount rate effect (French et al., 1987), despite the fact that a clear link between volatility and next period returns is weak and typically has the wrong sign.

Going forward, however, stock returns will be higher on average in the periods after the initial underreaction plays out. In fact, objective expected returns will remain high for longer than they would under rational expectations because the volatility shock effectively lasts longer in the agents minds. For example, even after (true) expected volatility has returned to normal the agent may believe volatility is high because they extrapolate from a period in which volatility had increased. Thus, the news about volatility will negatively forecast stock returns in the near term but positively forecast them in the longer term, despite the fact that news about volatility

strongly forecasts future volatility in the near term, but only weakly forecasts future volatility in the longer term. The variance risk premium will mirror this behavior with a drop in the near term followed by an increase in the longer term. Further, and importantly, the measured variance risk premium will forecast stock returns appropriately at virtually all horizons leading to a tight link between equity and variance risk premia as we see in the data (Bollerslev et al., 2009). However, in the model this occurs because of biased beliefs rather than because of rational risk premiums, and the model can account for the otherwise puzzling fact that while the variance risk premium is a strong forecaster of returns, neither the VIX or realized variance are individually strong forecasters of returns.

After developing the model and this intuition, we estimate the model parameters using GMM to quantitatively target the facts outlined above and show that we can do so reasonably well. Most importantly, our model nests the fully rational case but our parameter estimates (and associated standard errors) strongly support slow moving expectations about volatility. Further, we estimate the risk-return tradeoff parameters from the perspective of agents and find that there is a risk-return tradeoff in agents minds. However, their slow reaction to volatility causes this to be obscured in the data.

Next, we use survey data on volatility and uncertainty about stock returns from two sources (the Graham and Harvey CFO survey and the Shiller survey) and document that the surveys exhibit slow moving expectations as in our model. In particular, we regress survey expectations about volatility on past volatility realizations and show that expectations look like a weighted average of past volatility realizations as our model predicts, whereas optimal forecasts load only on recent volatility. This provides direct support for the mechanism of slow moving expectations in our paper.

We also document results at the firm level, where again we show implied volatility from firm level options does not react strongly enough to recent changes in volatility, leading to underreaction and a lower variance risk premium following increases in volatility (see also Poteshman (2001) for related work). This is true even when we include time fixed effects that control for aggregate movements in firm level volatility which makes a risk based explanation difficult since it focuses on purely idiosyncratic movements in firm level volatility. The firm level analysis provides further support for our story of underreaction and also provides robustness to our main empirical results which rely on aggregate market data and hence a relatively smaller sample.

Finally, we consider several potential objections to both the facts and our modeling choices. First, there is empirical evidence that volatility has two components: a higher frequency component and a lower frequency, more persistent component that our initial model ignores. It is possible that incorporating this component can match

some of the longer horizon empirical results: that expected returns and variance risk premiums do rise in the long run after variance shocks. However, this feature alone still fails to account for the short term behavior, and suggests that risk premia should be highest in the near term, counter to the data. Thus, just incorporating a long run component in variance but ignoring slow moving expectations will not easily match the patterns in the data. Second, we consider issues with extreme realizations of variance (which is positively skewed) in our main empirical facts. Third, in our baseline model volatility comes from the volatility of cash flows. This assumption is for simplicity but not crucial for our story. Fourth, we consider other explanations for our results including models with rational inattention and heterogeneous agents. While these models may indeed explain some features of the data, we explain why they do not easily generate the joint behavior of the facts we study. Finally, we study additional evidence of our channel including Nagel, Reck, Hoopes, Langetieg, Slemrod, and Stuart (2017) who show how investors respond to volatility changes empirically. In their data, more sophisticated and more experienced investors respond more quickly to changes in volatility. This makes sense if we expect these investors to have a smaller degree of bias in forming expectations of volatility.

1.1 Related Literature

Our model is similar to other models of extrapolation from past data including Barberis, Greenwood, Jin, and Shleifer (2015), Collin-Dufresne, Johannes, and Lochstoer (2016), and Nagel and Xu (2019), though our focus is on volatility rather than returns or cash flows.[6] Our paper also fits into a broader literature on under and overreaction, for example, Daniel, Hirshleifer, and Subrahmanyam (1998) and Barberis, Shleifer, and Vishny (1998). These papers are able to generate underreaction and delayed overreaction through potentially different underlying behavioral biases. At least qualitatively, this underreaction and delayed overreaction is precisely what is needed to match the data. We focus on extrapolation in particular to generate our results without taking a strong stand on where this extrapolation comes from at a deep level. We find extrapolation appealing both because it appears to exist in many contexts (e.g., Barberis et al. (2015)) and also because it is analytically very tractable in our setting. Further, there is extensive empirical evidence that agents do appear to extrapolate from past experiences when forming expectations (Malmendier and Nagel, 2011; Greenwood and Shleifer, 2014; Glaeser and Nathanson,

[6]The agents in our model also differ in spirit from other forms of extrapolation such as Nagel and Xu (2019) where agents rely too much on recent data because their memory "fades." In our context agents focus too little on recent data and too much on past data.

2017). Importantly, the survey data we study suggests extrapolation of volatility, hence suggests expectation formation similar to what we use in the model.

It may appear surprising that extrapolation in our context leads to *under*reaction as extrapolation is most often associated with *over*reaction. For example, in models where agents extrapolate from past returns or cash flows agents typically overreact to news (e.g., see Barberis et al. (2015)). However, this depends on the persistence of the true process and the noise about the conditional mean from observing past realizations. For returns or cash flow growth, whose conditional means move very slowly and are more difficult to measure using past returns or cash flows due to a low signal to noise ratio, it makes sense that extrapolating from recent data will lead to overreaction. Suppose the true mean of cash flow growth is constant, but investors take a moving average of the past year as an estimate. Then they will clearly overreact to new information when it arrives: they see current high cash flow growth and they update positively in a more aggressive manner than a rational Bayesian would. Our setting is different both because volatility is easier to observe and also because it moves very quickly – roughly speaking volatility follows an AR(1) with monthly autocorrelation around 0.75. Thus, the optimal forecast for future volatility is roughly todays volatility, and conditional on this there is little information in prior months. If one takes a moving average of past volatility that includes many months, then one would not put *enough* weight on recent volatility and put too much weight on past volatility. This leads to *under*reaction. The strength of this underreaction, and at which point it becomes overreaction, depends on how far back agents extrapolate. We feel this underreaction is quite natural in our setting – agents only extrapolate a modest amount compared to other settings using returns or cash flows.

In our model there is naturally underreaction in volatility expectations in the short term followed by delayed overreaction of expected volatility in the longer term. This matches the spirit of Giglio and Kelly (2017) who show focus on overreaction of long term volatility expectations. In particular, they argue that longer term expectations of volatility are too volatile relative to those at short horizons, a form of overreaction in long term expectations. However, we assume a simple structure for agents beliefs such that the dynamics for volatility under Q are easily captured by a simple AR structure – thus we can't speak directly to their result. Further, we have little to say about unconditional variance risk premia (Dew-Becker, Giglio, Le, and Rodriguez, 2017). We could obtain facts about unconditional variance risk premiums in extensions if agents price variance shocks or are biased on average, but in our baseline model agents beliefs are not biased on average, only conditionally. We focus on this case because we focus on matching facts related to the conditional,

rather than unconditional, behavior of risk premiums.

The fact that at the aggregate level changes in volatility negatively forecast stock returns while lags positively forecast mirror the results at the firm level documented by Rachwalski and Wen (2016).[7] In fact Rachwalski and Wen (2016) suggest a story similar to ours for their findings, though there are many important differences in our work. We focus on the aggregate market since this is where the standard risk-return relation in equilibrium asset pricing models should apply and we target the additional facts on the variance risk premium as well. We also study and estimate a quantitative model. Similarly, Poteshman (2001) documents underreaction and subsequent overreaction in options markets and attributes this to cognitive biases but does not link this to equity risk premiums and does not view this through the lens of a quantitative equilibrium model. We confirm and extend these firm level results

2. Stylized Empirical Facts

We begin by focusing on the main stylized empirical facts that we will target in our model. We study US data from 1990 to 2018 for which we have stock market excess returns, the VIX (taken as the VIX on the last day of the previous month, thus representing forward looking variance for the month) and realized variance (computed as the sum of squared daily returns within a month). Stock return data are from Ken French. In addition, we study variance swap returns and VIX futures returns from Dew-Becker et al. (2017) and Cheng (2018), respectively, though these have shorter samples (variance swap returns are 1996-2017 and VIX futures returns are 2004-2017). We define the variance risk premium (VRP) as the squared VIX minus realized variance, though to supplement this we also use the actual return series as well. When using returns (e.g., variance swap or VIX futures) we take the negative of the returns, so the implication is the return for selling variance or being short the VIX. This means the unconditional premium for both returns is positive as the exposure to volatility is negative.

Figure 1 Panel A plots impulse responses of realized variance (RV), excess returns, and the variance risk premium (VIX squared minus realized variance) to a shock to RV using a VAR with RV ordered first, so all variables can respond contemporaneously to a realized variance shock. We see RV follows roughly an AR(1),

[7] "Stocks with increases in idiosyncratic risk tend to earn low subsequent returns for a few months. However, high idiosyncratic risk stocks eventually earn persistently high returns." Rachwalski and Wen (2016)

spiking after the shock and mean reverting back after about 6 months (consistent with persistence of about 0.72 monthly). Stock returns fall contemporaneously with an increase in RV at time 0, consistent with a discount rate effect of volatility shocks (French et al., 1987). However, expected returns are negative in period 1 as well, meaning returns continue to fall next period on average. Expected returns continue to increase, eventually going positive several months out, but this is after the volatility shock has largely subsided from the perspective of the left panel. Thus risk premium rise later, after volatility has largely mean-reverted, which is quite different from a standard risk-return tradeoff view. The variance risk premium shows very similar patterns, with the predicted premium being negative in month 1 before slowly rising and becoming positive beyond month 3. Most notably, the premiums appear to rise as future volatility falls.

Panels B and C of Figure 1 study alternative specifications for this VAR. While Panel A does simple OLS, Panel B runs weighted least squares based on lagged stock market volatility, thus downweighting observations when volatility is expected to be high. We see largely the same patterns in Panel B as in Panel A. Panel C explores how influential high variance periods are for our conclusions. In particular, since realized variance is positively skewed, periods of extreme variance can be particularly influential. To explore this, we drop the highest 10% of observations based on ex-ante realized variance – much of these observations occur during the 2008-09 financial crisis period in our sample when realized variance was highest. Panel C shows roughly similar patterns, though it does suggest periods of high variance are important. In particular, the VRP response at time 1 is now muted, though the VRP continues to display the hump shaped pattern as before, as does the equity risk premium. Taken together, this supports a slow moving response of the premiums to realized variance, despite the fact that future realized variance still mean reverts fairly quickly. This is in contrast to the standard benchmark model, with the equity premium being affine in expected variance. In this setting the equity risk premium response should roughly mirror the response of future variance from period 1 onwards, with a spike upwards followed by a decline as future variance mean reverts.

The results are broadly consistent with the main empirical findings in the literature. First, the risk-return tradeoff overall is fairly weak, and often estimated to be negative, for stock returns on lagged realized variance (Glosten et al., 1993; Moreira and Muir, 2017, 2019). Relative to the literature we also show that the risk-return tradeoff appears to be flat or negative at first but increasing with horizon with a hump shaped response. In contrast, realized variance predicts realized variance strongly next month though this forecast fades with horizon (roughly consistent with variance following an AR(1) with monthly autocorrelation near 0.72).

This suggests that variance weakly or negatively predicts returns at horizons at which it strongly forecasts variance, but positively forecasts returns at longer horizons for which it only weakly forecasts variance. The time 0 response, which shows the return response to a contemporaneous shock to realized variance, is strongly negative as in French et al. (1987). Our variance risk premium results are consistent with this type of pattern as well. A variance shock initially predicts a decline in the variance risk premium (VIX squared minus realized variance) but then gradually predicts an increase in the variance risk premium. This result is consistent with Cheng (2018) who studies a claim on implied volatility (VIX) instead of realized variance. Finally, the variance risk premium and equity risk premium are tightly linked as in Bollerslev et al. (2009).

To further explore these facts and understand what drives the results, we run several additional forecasting regressions in Table 1. In particular, we compute changes in realized variance and realized volatility over the past six months and use this to forecast equity risk premiums, variance risk premiums, and future realized variance over the next month. Using the change focuses on whether the coefficients at the 1 and 6 month horizons are different from each other, versus testing whether they are (individually) different from zero. The results show that increases in variance over 6 months appear to negatively forecast returns next month (column 1) and negatively forecast variance risk premiums (columns 5-7), but positively forecast future variance (column 9). Importantly, the table uses actual returns on variance claims rather than the implied premium inherent in taking squared VIX minus future realized variance as used in the VAR. Columns 6 and 7 use variance swap returns Dew-Becker et al. (2017) and VIX futures returns Cheng (2018) and show they are predictable by changes in variance. The negative sign indicates that it is cheap to insure against future increases in volatility when past volatility increases, similar to the results in Cheng (2018). Column 8 predicts future realized variance using the 1 and 6 month lags of realized variance individually, and confirms that the bulk of the information in expected variance comes from the first lag, consistent with variance following an AR(1) with autocorrelation of around 0.72.

We repeat this analysis in Panel B using volatility in place of variance, which will have fewer extreme realizations. The variance risk premium return results hold, with increases in volatility predicting negative premiums. Volatility also predicts future volatility with about the same pattern shown before for variances. However, now the six month change in volatility only weakly negatively predicts future returns, highlighting that much of the results in column 1 of Panel A come from high variance realizations. While this suggests caution in terms of reliably predicting stock returns with changes in variance, our main point is in the benchmark models this coefficient

would typically be strongly positive rather than negative, which is at odds with the data.

In column 2 we replicate the results from Bollerslev et al. (2009) that the variance risk premium is a robust predictor of stock returns, and this result is even stronger in our sample which adds the more recent ten years of data compared to the sample used in Bollerslev et al. (2009). Thus, there is a strong link between implied variance risk premiums and equity risk premiums. Importantly, though, the variance risk premium is a strong predictor of returns, though the VIX or realized variance individually are not as show in columns 4 and 5.[8] VIX squared has little to no forecasting power for returns, and if anything has the wrong sign with a negative coefficient. Realized variance appears to predict returns weakly, but also with a negative sign. This is puzzling from the perspective of the model in Bollerslev et al. (2009) in which the VIX alone is a strong predictor of returns, both because it embeds the variance risk premium and because it reflected expected future variance, and both of these strongly contributed to the equity risk premium. These results continue to hold in Panel B when using volatility in place of variance.

Standard asset pricing models struggle with these facts because they typically suggest that an increase in risk (volatility) will be associated with heightened risk premiums at all horizons, and this relationship will be strongest in the near term and will decay with horizon. For example, Moreira and Muir (2017) show the risk return tradeoff in leading models is strong, including models with habit formation (Campbell and Cochrane, 1999), long run risk (Bansal and Yaron, 2004; Drechsler and Yaron, 2011), rare disasters (Barro, 2006; Wachter, 2013) and intermediary models (He and Krishnamurthy, 2013). Further, expected returns will typically rise most on impact and will gradually fade through time as volatility fades. We are not aware of leading equilibrium asset pricing models which produce a temporary decline in risk premiums followed by a delayed increase.

Table 2, which we return to later, repeats this analysis at the firm level for US data which gives us significantly more observations compared to the aggregate results and hence provides robustness to our main results. We see strikingly similar results for the variance risk premium with increases in variance negatively predicting the variance risk premium. This is true even when including time fixed effects suggesting that the facts we document hold even when removing aggregate movements in volatility.

We also find supportive evidence in survey data: surveys that captures investors perception of volatility or uncertainty are slow moving and load significantly more on past realizations of volatility compared to optimal forecasts. We return to the survey evidence in a later section.

[8]See also **?** who construct a historical time series meant to track uncertainty.

We return to these facts after presenting the model, and we discuss additional empirical robustness in the Appendix, including results using international data. In the Appendix, we show that the US data of returns on realized variance is robust going back to the 1950s but changes if one includes the Great Depression (see Figure 8). The Great Depression era has particularly large values for realized variance which dominate the regression. However, most importantly, even including this data there is no strong evidence of a risk-return tradeoff (e.g., variance does not reliably predict future returns) though the evidence of a statistically significant negative coefficient on the next month return changes when including this period. The point estimate is still negative but not significant, though again our main point is that it is not strongly positive as most models would predict. Hence, we interpret the negative one month prediction with some caution in light of the longer sample. In our parameter estimation, we focus on the more recent US data from 1990 when we have the VIX and variance risk premium data.

3. The Model

In this section, we present and estimate an asset pricing model similar to that in Bollerslev, Tauchen and Zhou (2009) except that we allow the representative investor to have biased beliefs regarding the dynamics of stock return volatility. This simple modification enables the model to account for the empirical evidence discussed earlier.

Let aggregate log dividend growth be given by:

$$\begin{aligned} \Delta d_t &= \mu + \sigma_t \varepsilon_t, & (2) \\ \sigma_t^2 &= \bar{v} + \rho\left(\sigma_{t-1}^2 - \bar{v}\right) + \omega\eta_t, & (3) \end{aligned}$$

where σ_t^2 is the realized variance of dividend growth innovations, observed at time t, and ε_t and η_t are uncorrelated i.i.d. standard Normal shocks. Variance is persistent with $0 < \rho < 1$.[9]

We assume a representative stockholder with consumption equal to aggregate dividends and whose marginal utility prices all claims in the economy. The agent's

[9]Bollerslev, Tauchen, and Zhou (2009) additionally let the variance of variance follow a square root process. In the appendix, we show that this simplification is unimportant for our conclusions. Equation (3) implies that variance can go negative. For ease of exposition we follow, e.g., Bansal and Yaron (2004) and Bollerslev, Tauchen, and Zhou (2009), and proceed as if σ_t^2 is always non-negative.

expectations of the conditional variance of dividend growth are given by:

$$\begin{aligned} E_{t-1}^{S}\left[\sigma_t^2\right] &= \bar{v} + \rho x_{t-1}, \quad (4) \\ x_t &= \phi x_{t-1} + (1-\phi)\left(\sigma_t^2 - \bar{v}\right) \\ &= (1-\phi)\sum_{j=0}^{\infty}\phi^j\left(\sigma_{t-j}^2 - \bar{v}\right), \quad (5) \end{aligned}$$

where $0 \leq \phi < 1$. The S superscript on the expectations operator highlights that the expectation is taken under the agent's *subjective* beliefs. If $\phi = 0$, the agent has rational expectations about the volatility dynamics, while if $\phi > 0$ the agent has slow-moving volatility expectations, allowing an exponentially weighted average of past variance to affect the current expectation, as opposed to only the current value as the physical volatility dynamics prescribe.

Under agents' beliefs, the shock to variance is:

$$\begin{aligned} \omega\eta_t^S &\equiv \sigma_t^2 - \bar{v} - \rho x_{t-1} \\ &= \rho\left[\left(\sigma_{t-1}^2 - \bar{v}\right) - x_{t-1}\right] + \omega\eta_t, \quad (6) \end{aligned}$$

where $\rho\left[\left(\sigma_{t-1}^2 - \bar{v}\right) - x_{t-1}\right]$ is the mistake agents make when forecasting variance. We can thus write the dynamics of x_t under agents' beliefs as:

$$x_t = \left(\phi + (1-\phi)\rho\right)x_{t-1} + (1-\phi)\omega\eta_t^S. \quad (7)$$

Note that investors' variance expectations are sticky relative to the true variance dynamics if $\phi > 0$, as in this case $\phi + (1-\phi)\rho > \rho$. Also note that the shock itself is moderated by a factor of $(1-\phi)$. The top plot of Figure 2 shows the impulse-response from a positive variance shock (η_0) for objective and subjective expected variance. The parameter values are as estimated in the data in the estimation section below. The true AR(1) dynamics of variance are reflected in the monotonically decaying response of rational case (dashed red lines). The solid blue lines give the impulse-response of agents' expected variance, as reflected by the dynamics of x_t. Agents' initially underreact, as $\phi > 0$ in this case, but the higher persistence of x_t leads to subsequent overreaction.

Following Bollerslev, Tauchen, and Zhou (2009), the agent has Epstein-Zin utility (Epstein and Zin, 1989) where β, γ, and ψ are the time-discounting, risk aversion, and intertemporal substitution parameters, respectively. The stochastic discount factor is therefore:

$$M_t = \beta^{\theta} e^{-\frac{\theta}{\psi}\Delta d_t + (\theta-1)r_t}, \quad (8)$$

where $\theta = \frac{1-\gamma}{1-1/\psi}$ and r_t is the log return to the aggregate dividend claim. We use the standard log-linearization techniques of Campbell and Shiller (1988) and Bansal

and Yaron (2004) to derive equilibrium asset prices (see Appendix for details). In particular, we assume aggregate log returns are $r_t = \kappa_0 + \kappa pd_t - pd_{t-1} + \Delta d_t$, where pd is the aggregate log price-dividend ratio and κ is a constant close to but less than one that arises from the log-linearization. We then obtain:

$$pd_t = c - Ax_t, \tag{9}$$

where $A = -\frac{1}{2}\frac{\rho(1-\gamma)(1-1/\psi)}{1-\kappa(\phi+\rho(1-\phi))}$. Notice that if $\gamma, \psi > 1$ we have that $A > 0$ and the price-dividend ratio is low when agents perceive variance to be high, as in the data. This is the standard preference parameter configuration for asset pricing models with Epstein-Zin preferences.

3.1 Equity risk premium dynamics

Let r_t and $r_{f,t}$ denote the aggregate log return and risk-free rate in period t, respectively. The subjective risk premium of log returns can be written:

$$E^S_{t-1}\left[r_t - r_{f,t}\right] = \gamma E^S_{t-1}\left[\sigma^2_t\right] + (1-\theta)\,\Theta - \frac{1}{2}Var^S_{t-1}\left(r_t\right), \tag{10}$$

where $\Theta = \left(\rho A\left(1-\phi\right)\omega\right)^2$ captures the price effect of discount rate shocks due to the variance shocks. The first term is a standard risk-return trade-off that is linear in the conditional variance of dividend growth. The second constant term is due to the persistent discount rate shock and the Epstein-Zin preferences, where persistent shocks are priced if $\theta \neq 1$. This term is constant as shocks to variance are homoscedastic. The last term is a Jensen's inequality term that arises as we are looking at the risk premium of log returns. The conditional variance of log returns is determined both by the conditional variance of dividend growth and the impact of the variance shock on the price-dividend ratio:

$$Var^S_{t-1}\left(r_t\right) = \Theta + E^S_{t-1}\left[\sigma^2_t\right]. \tag{11}$$

The *objective* risk premium, however, is:

$$E^P_{t-1}\left[r_t - r_{f,t}\right] = E^S_{t-1}\left[r_t - r_{f,t}\right] - \kappa\left(1-\phi\right)A\left(E^P_{t-1}\left[\sigma^2_t\right] - E^S_{t-1}\left[\sigma^2_t\right]\right), \tag{12}$$

where the P superscript on the expectation denotes that it is taken using the true, objective variance dynamics. If $\phi > 0$, agents make mistakes in their conditional variance expectations. These mistakes are reflected in current discount rates and therefore prices. Consider a positive shock to variance ($\eta_{t-1} > 0$). With $\phi > 0$

investors' expectations are sticky, meaning investors do not update their beliefs sufficiently and initially underreact to the variance shock. Thus, $E^P_{t-1}[\sigma^2_t] > E^S_{t-1}[\sigma^2_t]$. Since $A > 0$ in the relevant calibrations, this means a *positive* shock to variance can, if the mistake is sufficiently large, *decrease* the objective risk premium. The reason is that investors will on average perceive a positive shock to discount rates next period as the realized value of σ^2_t on average is higher than they had expected. This leads to a predictable decline in the price-dividend ratio under the objective measure. The bottom plot of Figure 2 shows the impulse-response of the log price-dividend ratio to a volatility shock. As expect the price-dividend ratio falls at the impulse, but note that it keeps falling for two more quarters due to the increase in discount rates when agents learn variance is higher than expected. Subsequently, given the too persistent variance expectations, agents eventually overreact to the volatility shock, which leads to $E^P_{t+j-1}\left[\sigma^2_{t+j}\right] < E^S_{t+j-1}\left[\sigma^2_{t+j}\right]$ for some $j > 0$. In this case, the second term in Equation (12) becomes positive and the conditional risk premium overshoots.

3.2 Variance risk premium dynamics

In addition to the equity claim, we also price a variance claim with payoff:

$$RV_t \equiv \Theta + \sigma^2_t, \tag{13}$$

where RV_t stands for realized variance at time t. We define the time $t-1$ implied variance (IV_{t-1}) as the swap rate that gives a one-period variance swap a present value of zero:

$$0 = E^S_{t-1}\left[M_t\left(RV_t - IV_{t-1}\right)\right]. \tag{14}$$

Thus:

$$IV_{t-1} = E^S_{t-1}\left[R_{f,t}M_t RV_t\right]. \tag{15}$$

As is standard in the literature, we denote the (objective) expected payoff of a position in the variance swap where you are paying the realized variance and receiving the implied variance as the variance risk premium:

$$VRP_{t-1} = IV_{t-1} - E^P_{t-1}\left[RV_t\right]. \tag{16}$$

If realized variance is high (low) in bad times, this risk premium is positive (negative).

Our model-definition of realized variance is motivated by industry practice for variance swap payoffs, where monthly realized variance is the sum of squared daily log returns within the month. In the model, squared monthly log returns are:

$$\left(r_t - E_{t-1}\left[r_t\right]\right)^2 = \Theta\eta^2_t + 2\sigma_t\Theta^{0.5}\eta_t\varepsilon_t + \sigma^2_t\varepsilon^2_t. \tag{17}$$

To approximate the use of higher frequency data to estimate realized variance within our model, we assume that the second moments of realized shocks equal their continuous-time limit.[10] Setting $\eta_t^2 = \varepsilon_t^2 = 1$ and $\eta_t \varepsilon_t = 0$ in Equation (17) gives the realized variance in Equation (13).[11]

The equilibrium implied variance is:

$$IV_{t-1} = E_{t-1}^S [RV_t] + k, \tag{18}$$

where $E_{t-1}^S [RV_t] = E_{t-1}^S [\Theta + \sigma_t^2] = \Theta + \bar{v} + \rho x_{t-1}$ and $k = \left(\frac{1}{2}\gamma^2 - \frac{1/\psi - \gamma}{1 - 1/\psi} \kappa (1 - \phi) A \right) \omega^2$. The second term is an unconditional risk premium required by the agents due to the variance claim's exposure to shocks to variance. The conditional variance risk premium is then:

$$\begin{aligned} VRP_{t-1} &= IV_{t-1} - E_{t-1}^P [RV_t] \\ &= k + E_{t-1}^S [RV_t] - E_{t-1}^P [RV_t] \\ &= k + E_{t-1}^S \left[\sigma_t^2\right] - E_{t-1}^P \left[\sigma_t^2\right]. \end{aligned} \tag{19}$$

Thus, the dynamics of the variance risk premium share a component of the dynamics of the equity risk premium (Equation (12)), namely the mistakes agents' make in their variance expectation. Thus, agents will initially underreact to the variance shock, but subsequently overreact due to their sticky expectations, which leads to time-variation in the variance risk premium similar to that in the data. In fact, the lagged variance risk premium forecasts equity returns, as it does in the data and as it does in Bollerslev, Tauchen, and Zhou (2009). However, in the their model this is due to time-varying variance of variance, which we abstract from in this baseline version of our model.

Next, we estimate the parameters of the model to assess if it can quantitatively account for the empirical observations discussed earlier.

[10] That is, if $W_t^{(1)}$ and $W_t^{(2)}$ are standard Brownian motions with uncorrelated innovations, $\int_t^{t+1} \left(dW_t^{(j)} \right)^2 = \int_t^{t+1} dt = 1$ for $j = \{1, 2\}$ and $\int_t^{t+1} dW_t^{(1)} dW_t^{(2)} = 0$.

[11] In benchmark equilibrium models, typically calibrated at the monthly frequency (e.g., Bollerslev, Tauchen, and Zhou (2009), Drechsler and Yaron (2011)), there is no clear counterpart to this multi-frequency approach where IV and RV are monthly, but where RV is estimated using daily data. In the models cited above, the definition of the IV_t is the risk-neutral expectation of the market return variance in month $t + 2$. For example, IV at the end of January is the risk-neutral expectation at the end of January of market return variance in March. We define RV in a manner that avoids this one-month offset that is at odds with the data definitions. This brings the model closer to the moments from the data we use for estimation of the model parameters. While it is convenient to align the model definitions more closely to the timings used in the data, we note that our model results would also go through with alternate definitions of the variance risk premium used in earlier literature.

3.3 Model estimation

We estimate the model using the Generalized Method of Moments (GMM). The data is monthly and from 1990 through 2018. We use the VIX_t^2 as the proxy for IV_t, where VIX_t is the option-implied risk-neutral volatility of stock returns over the next month. RV_t is calculated as the sum of daily squared log excess market returns through in month t. We choose moments that are at the heart of the issues we seek to address with the model – including the unconditional variance risk premium and the main predictive relations between variance and returns. In particular, the moment conditions are: the mean, variance, and first autocovariance of RV; the covariance of shocks to RV with contemporaneous log excess market returns, the covariance of RV and next month's log excess market returns, the covariance of IV_t and next month's log excess market returns; the mean of the realized variance risk premium, $IV_t - RV_{t+1}$; the mean of log excess returns; the covariance of RV_t with next month's realized variance risk premium, $VIX_t^2 - RV_{t+1}$.

We run a two-stage efficient GMM. The parameter vector is $\Phi = \{\bar{v}, \rho, \omega, \gamma, \psi, \phi\}$, where $\kappa = 0.97^{1/12}$. Note that our moments do not require us to estimate the time-discounting parameter, β. The first-stage weighting matrix, W, is a 9×9 diagonal matrix, with the inverse of the variance of each moment conditions' exogenous components along the diagonal. The second stage weighting matrix is the usual inverse of the spectral density matrix, where the latter is calculated using Newey-West with 12 lags and the first stage parameter estimates.

Table 4 shows the estimated model parameters along with their standard errors. The degree of extrapolation (ϕ) is 0.66 with a standard error of 0.04. Thus, we strongly reject rational expectations within our model (the case where $\phi = 0$). Risk aversion γ is 5.99 (standard error 0.85) while the elasticity of substitution ψ is 2.87 (standard error 1.12). Thus, the agent has a preference for early resolution of uncertainty, as in the Bollerslev, Tauchen, and Zhou (2009) model, which implies she is averse to persistent shocks to the distribution of future consumption. In our model this means that variance risk is priced and that the unconditional variance risk premium is positive, as in the data.

Figure 4 shows impulse responses from a VAR with RV, log excess market returns, and realized VRP, as in the data analysis of Section 2. The shocks are ordered in the same manner. The model implies a VAR(1), and we run a VAR(1) also in the data to compare with the model. The impulse-responses in the data are plotted with a black solid line, with the usual two standard error bands. The impulse-responses from the estimated model are given in the blue dashed lines, while the impulse-responses from the model assuming $\phi = 0$ (rational case) are given in the red dash-dotted lines.

Both models are consistent with the autocorrelation pattern of RV. A positive

shock increases variance on impact and decays monotonically and relatively quickly. The impact of an RV shock on the conditional market and variance risk premiums is very different across the two models. In particular, while in both the rational and extrapolative models market returns decrease contemporaneously with a positive shock to variance, the response in the rational model is to immediately increase the conditional risk premium due to the usual risk-return trade-off (Equation (12)), at odds with the empirical facts. In the extrapolative model, however, the response of the objective conditional equity premium is negative the first two months due to the mistake investors are making in their variance forecast, as shown in Figure 2. The equity premium subsequently overshoots due to the slow-moving expectations of the extrapolative agents, consistent with the pattern in the data. The same is true for the variance risk premium, although in this case the pattern is stronger as its dynamics are only affected by the mistake in expectations (see Equation (19)). The rational model has no effect on the variance risk premium from an RV shock, again at odds with the data.[12]

4. Additional Evidence: Survey Data and Firm Level Analysis

4.1 Survey Data on Volatility

Survey data on investors expectations is especially useful because it allows us to evaluate the main mechanism in our model using direct data on expectations. In addition, Giglio, Maggiori, Stroebel, and Utkus (2019) show that survey data on investor beliefs about risk translates directly into actions in terms of portfolio allocations. Specifically, they find that investors substantially reduce their portfolio allocation to stocks when they think stocks are riskier in terms of greater probability of a significant decline in the stock market.

We bring survey data related to volatility from two sources. The first is the Graham and Harvey survey of CFOs which is quarterly from 2001. The survey asks respondents for a mean forecast for the stock market over the next year as well as 10th and 90th percentiles. We construct the 90th minus 10th percentile as a measure of volatility or uncertainty and square this number to get a measure of

[12]This is true also in the model of Bollerslev, Tauchen, and Zhou (2009), even though they do have time-variation in the expected variance risk premium. This occurs as in their model shocks to the variance of variance are independent of shocks to variance.

expected variance. While this measure has limitations, it does capture how spread out agents view the return distribution, and under the view of a normal distribution would perfectly capture agents expectations about volatility. Our second source of survey evidence is from Robert Shiller who asks investors the probability of a stock market crash over the next 6 months such as that seen in 1987. Again, we view this as correlated with volatility though it is still an imperfect measure. We use the monthly Shiller sample which begins in July of 2001.

We proceed in two ways to assess whether survey data display slow moving expectations about volatility. First, we fit survey expectations and actual realized variance over the period investors are asked to forecast as an exponential weighted average of past variance:

$$y_{t\to t+k} = a + b\sum_{i=1}^{J} \phi^{i-1}\sigma^2_{t-i} + \varepsilon_t$$

where $y_{t\to t+k}$ is the actual future realized variance from time t to $t+k$, and then $y_{t\to t+k}$ is replaced by survey expectations of variance instead. When using realized variance on the left hand side, we compute forward looking realized variance over the horizon which corresponds to the survey expectations about volatility (e.g., k is 1 year for CFO survey and 6 months for the Shiller survey). We take J, the number of lags of realized variance, to correspond to 12 months in the monthly Shiller survey, and 6 quarters in the quarterly survey (longer lags produce similar results). We estimate ϕ in both cases, where a higher ϕ from survey data indicates more reliance on past variance compared to the optimal forecast. This specification has the benefit that it maps exactly to our model setup for beliefs.

Table 3 gives our estimates. We find $\phi_{survey} > \phi_{RV}$ meaning survey expectations depend much more on past volatility than optimal forecasts indicate. We find ϕ_{RV} is economically and statistically fairly small, consistent with realized variance being fairly well approximated by an AR(1), while we find ϕ_{survey} to be large, between 0.75 (Shiller) and 0.85 (CFO). The estimates from these two totally separate surveys thus deliver similar degrees of extrapolation from past variance. The substantial amount of extrapolation we see in the survey data is a bit larger than what we estimate from our model, meaning the expectations bias from pure survey data appears stronger than the bias we estimate from only financial market data.

We plot the implied loadings on each of the lags of volatility (where implied loadings are normalized to sum to one) based on the estimated ϕ in Figure 4. The patterns are similar in both cases: the actual volatility process loads mostly on the first lag of volatility, and very little on subsequent lags. In contrast, the survey

expectations show large dependence of volatility far into the past (high ϕ). This provides independent evidence consistent with our model.

To further assess the degree of investors slow moving expectations, we run a vector autoregression (VAR) with future realized variance as well as the reported expectations from the survey. We order future variance first, followed by the survey expectation and plot the impulse response to a variance shock. Results are given in Figure 5. As before, future variance increases substantially after this shock then subsequently declines as it mean reverts. The survey expectations, however, show a hump shaped response, consistent with expectations continuing to rise after the initial shock. The expectations initially do not rise as much (underreaction) but then subsequently remain elevated long after expected variance declines (subsequent overreaction), consistent with the dependence of survey expectation on longer lags of past variance. The pattern from both surveys is similar, and this prediction is exactly what we expect from our model of slow moving expectations of volatility. Thus, two independent surveys provide consistent evidence in favor of the mechanism we propose.

4.2 Firm Level Analysis

We revisit our stylized aggregate facts at the firm level (stock level). We take implied volatility from OptionMetrics at the stock level from 1996-2017 and use daily and monthly return data from CRSP for the stocks in the merged OptionMetrics sample (6,489 unique stocks over the sample). Implied volatility is measured on the last day of the month and measures option implied volatility over the subsequent month (30 days) for at the money options. Realized variance is computed using the daily returns within a given month. Our measure of the variance risk premium is then $IV_{i,t}^2 - RV_{t+1}$ which is the implied variance over the next month minus the actual realized variance over the next month. We use daily log stock returns from CRSP and computed the sum of squared log returns over the following month's trading days as our measure of realized variance.

Analogous to our results in Table 1, we forecast equity risk premiums, variance risk premiums, and future realized variance over the next month using the change in realized variance from month t to $t-6$. We winsorize lags of realized variance at the 90th percentile, though importantly we do not winsorize future realized variance so that the left hand side in this case is still the realized variance risk premium. In unreported results we find similar result without winsorization, but the main advantage is we find much stronger predictive power for future variance with winsorization due to substantially more noise in firm level realized volatility estimates compared to the

aggregate. We also find qualitatively similar results in several other specifications, including using log of realized variance or using volatility in place of variance, though these results are omitted for space.

The results show that increases in volatility over 6 months negatively forecast variance risk premiums, but positively forecast future variance. The coefficients for predicting future variance and future variance risk premiums are highly statistically significant with or without time fixed effects (standard errors are double clustered by time and firm). The results with time fixed effects are especially important because these remove any aggregate movements in firm level variance or variance risk premiums. By removing aggregate effects, we are more likely capturing purely idiosyncratic movements in realized variance that helps push against a risk-based story for our results. These results are also similar in spirit to Poteshman (2001) who argues for underreaction in option prices in an earlier sample.

The firm level analysis achieves two things. First, it provides robustness to our aggregate results which rely on fewer observations. Second, it provides more insight into whether the variance risk premium results we document are likely driven by true economic risk premiums (compensation for risk) or whether they are instead more likely driven by biased expectations and underreaction to changes in volatility. As stressed earlier, the aggregate results are not consistent with standard risk based models since higher risk (more variance) should, if anything, imply a higher rather than lower risk premium. Nevertheless, it is always possible to construct a model in which investor preferences move in such a way to match the aggregate evidence. The firm level evidence is more powerful since we think of firm level variance as largely idiosyncratic, especially in our second specification where we include time fixed effects in the regression to remove any common components of firm level variance. Hence, we would likely expect a much smaller effect at the firm level from a risk premium story due to variance shocks being more idiosyncratic at the firm level. Instead, we recover a coefficient of around -0.1 for the firm level VRP, which is very consistent with the magnitudes we observe in the aggregate results.

Further in this dimension, at the firm level we see a weakly negative but not significant coefficient for the equity risk premium. This is exactly what we expect in the model if agents do not price idiosyncratic firm level risk. In our aggregate results, investors should require more compensation for the increase in variance and this mechanism combined with biased beliefs results in the negative coefficient on the equity risk premium. Absent this channel, we would only expect the results to hold for the variance risk premium. Taken together, the firm level results strongly support our main hypothesis that agents initially underreact to changes in variance and that this is reflected in implied volatilities.

4.3 Evidence on Actual Trading Behavior

Nagel et al. (2017) show evidence that investors do react to changes in volatility with more sophisticated investors and older investors responding more strongly. Specifically, they show that higher income and older investors sell more aggressively following increases in volatility. This is reasonable in our model if one takes higher income investors to be more sophisticated and less prone to the expectations bias in our paper. Similarly, it is possible that investors learn more about the volatility process with time (as the evidence on investor experience suggests they would) and hence exhibit less of a bias as they are older. A shortcoming of our model is that it features a representative investor and so does not speak directly to this evidence (as there is no trade in equilibrium), though modest extensions of the model which allow for differences in the amount of bias would naturally be consistent with the evidence on trading behavior.

5. Extensions and Alternative Explanations

5.1 Model Extensions

We extend the model to incorporate richer, and more realistic, volatility dynamics. In particular we assume that the volatility of volatility is time varying along the lines of Bollerslev et al. (2009). This helps us match variation in the volatility of volatility. An appendix discusses this extension.

In the main model we put the stochastic volatility on the cash flow process. However, this is not particularly important and our paper doesn't have much to say whether this is discount rate or cash flow volatility. In our model discount rate volatility would still be priced and would still imply the highest premium at shorter horizons. We make the assumption of stochastic cash flow volatility for convenience.

Our model has implications for the price dividend ratio that are clearly rejected in the data. Most importantly the model – if taken literally – says the dividend yield is perfectly correlated with the VIX, which is clearly counterfactual. In particular, empirically the dividend yield is much more persistent than the VIX, though the two are positively correlated. This highlights that the dividend yield could also be driven by forces outside our model. In particular, an extension of our model with time-varying expected dividend growth would generate additional movements in the dividend yield and, if they were highly persistent, could generate the difference in persistence. This highlights why we choose not to use dividend yield related moments in our GMM estimation even though our baseline model has implications for these

moments.

5.2 Alternative Explanations

Moreira and Muir (2017) show that leading equilibrium asset pricing models (e.g., habits models, intermediary models, long run risk, and rare disasters) typically imply a strong risk return tradeoff and hence won't match the facts that volatility is a weak predictor of returns.

What other models could explain our results? While some models can indeed match some of our stylized facts, we are not aware of models that can quantitatively jointly match them. This is especially true regarding the firm level analysis which relies solely on idiosyncratic movement in firm level variance, and our survey expectation data which suggests slow moving volatility expectations. We briefly discuss models with rational inattention and heterogeneity in terms of which facts they can explain.

Models featuring infrequent rebalancing and/or rational inattention (Abel, Eberly, and Panageas, 2013) at first appear promising but won't easily match the facts that we document. Essentially, even if a small fraction of traders is attentive at any given time, they will still price in changes to volatility. Similarly, even agents know they will not rebalance again soon they will still ensure a risk-return tradeoff at the horizon at which they expect rebalance. This will result in a risk-return tradeoff that resembles the standard case. Further, Nagel et al. (2017) show evidence that investors do react to changes in volatility with more sophisticated investors (e.g., those in highest income brackets) responding most quickly. That is, it does not appear agents are not aware and do not act on changes in volatility. Finally, we are unaware of these models being able to easily match the variance risk premium dynamics, and particularly the firm level facts or the survey expectation data.

Heterogenous agents models can potentially explain the weak risk-return relation, and in these models this risk-return relation can even go negative depending on the wealth distribution (e.g., Gârleanu and Panageas (2015), Longstaff and Wang (2012)). These models feature a conditional risk-return tradeoff that is typically positive for most parts of the state space but can turn negative in the worst states. For the unconditional risk-return tradeoff to be weak, calibrations of the models would typically also require that the correlation between contemporaneous returns and volatility would be weak, which is not the case. It is not obvious these models would be able to explain the mismatch in frequencies that we observe, e.g., with risk premiums initially declining but then rising further out after volatility increases. Further, it is less clear that these models can match the variance risk premium results,

the firm level results we document (which rely on firm level idiosyncratic variance rather than aggregate variance), and the slow moving expectations from our survey data. In these models the relation between volatility and expected returns is only weak or negative in bad times though we don't find such a conditional relation in the data (for example, the strong negative correlation of returns and realized variance innovations is robust in good and bad market conditions).

6. Conclusion

We show that underreaction followed by delayed overreaction can match many empirical facts surrounding volatility and risk premiums that are puzzling from leading equilibrium asset pricing models. We achieve this feature by assuming agents extrapolate from past volatility and we estimate the degree to which they do so in the data. In particular, our model matches the weak overall risk-return tradeoff and matches the dynamic responses of both the equity premium and variance risk premium following shocks to variance. Survey evidence directly supports the idea that agents have slow moving expectations about volatility, as does evidence at the firm level.

References

Abel, Andrew B, Janice C Eberly, and Stavros Panageas, 2013, Optimal inattention to the stock market with information costs and transactions costs, *Econometrica* 81, 1455–1481.

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

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer, 2015, X-capm: An extrapolative capital asset pricing model, *Journal of Financial Economics* 115, 1 – 24.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics* 49, 307 – 343.

Barro, Robert J., 2006, Rare disasters and asset markets in the twentieth century, *The Quarterly Journal of Economics* 823–866.

Bollerslev, Tim, George Tauchen, and Hao Zhou, 2009, Expected Stock Returns and Variance Risk Premia, *The Review of Financial Studies* 22, 4463–4492.

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

Cheng, Ing-Haw, 2018, The VIX Premium, *The Review of Financial Studies* 32, 180–227.

Collin-Dufresne, Pierre, Michael Johannes, and Lars A. Lochstoer, 2016, Asset Pricing When 'This Time Is Different', *The Review of Financial Studies* 30, 505–535.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, 1998, Investor psychology and security market under- and overreactions, *The Journal of Finance* 53, 1839–1885.

Dew-Becker, Ian, Stefano Giglio, Anh Le, and Marius Rodriguez, 2017, The price of variance risk, *Journal of Financial Economics* 123, 225 – 250.

Drechsler, Itamar, and Amir Yaron, 2011, What's vol got to do with it, *Review of Financial Studies* 24, 1–45.

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

Gârleanu, Nicolae, and Stavros Panageas, 2015, Young, old, conservative, and bold: The implications of heterogeneity and finite lives for asset pricing, *Journal of Political Economy* 123, 670–685.

Giglio, Stefano, and Bryan Kelly, 2017, Excess Volatility: Beyond Discount Rates*, *The Quarterly Journal of Economics* 133, 71–127.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus, 2019, Five facts about beliefs and portfolios, Working paper, National Bureau of Economic Research.

Glaeser, Edward L., and Charles G. Nathanson, 2017, An extrapolative model of house price dynamics, *Journal of Financial Economics* 126, 147 – 170.

Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle, 1993, On the relation between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance* 48, 1779–1801.

Greenwood, Robin, and Andrei Shleifer, 2014, Expectations of Returns and Expected Returns, *The Review of Financial Studies* 27, 714–746.

He, Zhiguo, and Arvind Krishnamurthy, 2013, Intermediary asset pricing, *The American Economic Review* 103, 732–770.

Longstaff, Francis A., and Jiang Wang, 2012, Asset Pricing and the Credit Market, *The Review of Financial Studies* 25, 3169–3215.

Malmendier, Ulrike, and Stefan Nagel, 2011, Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?*, *The Quarterly Journal of Economics* 126, 373–416.

Martin, Ian, 2016, What is the expected return on the market?, *Quarterly Journal of Economics* 132(1), 367–433.

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

Moreira, Alan, and Tyler Muir, 2017, Volatility-managed portfolios, *The Journal of Finance* 72, 1611–1644.

Moreira, Alan, and Tyler Muir, 2019, Should long-term investors time volatility?, *Journal of Financial Economics* 131, 507 – 527.

Nagel, Stefan, Daniel Reck, Jeffrey Hoopes, Patrick Langetieg, Joel Slemrod, and Bryan Stuart, 2017, Who sold during the crash of 2008-9? evidence from tax return data on daily sales of stock, Working paper, National Bureau of Economic Research.

Nagel, Stefan, and Zhengyang Xu, 2019, Asset pricing with fading memory, Working paper, University of Chicago.

Poteshman, Allen M., 2001, Underreaction, overreaction, and increasing misreaction to information in the options market, *The Journal of Finance* 56, 851–876.

Rachwalski, Mark, and Quan Wen, 2016, Idiosyncratic Risk Innovations and the Idiosyncratic Risk-Return Relation, *The Review of Asset Pricing Studies* 6, 303–328.

Wachter, Jessica A., 2013, Can time-varying risk of rare disasters explain aggregate stock market volatility?, *Journal of Finance* 68, 987–1035.

7. Tables / Figures

Table 1: Stylized Facts. We run predictive regressions of future excess stock returns (market returns over the risk free rate), future variance risk premiums (measured using $VIX^2 - RV$, variance swap returns r_{var}, and VIX futures returns r_{VIX}), and future realized variance on various measures of past variance, change in past variance over 6 months ($\Delta_6\sigma_t^2$), and implied volatility from the VIX. In our notation σ_t^2 represents the realized variance of daily market returns in month t. The returns on variance swaps and VIX futures have a negative sign, thus representing the premium for insuring against future increases in VIX or variance (so that the variance risk premium is positive on average). Data are monthly from 1990-2018, the variance swap and VIX futures data are 1996-2017 and 2004-2017, respectively. Standard errors in parentheses use Newey West correction with 12 lags.

Panel A: Variance

	Excess Stock Returns				Variance Risk Premium			Variance	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	$r^e_{M,t+1}$	$r^e_{M,t+1}$	$r^e_{M,t+1}$	$r^e_{M,t+1}$	$VIX^2_t - \sigma^2_{t+1}$	$r_{var,t+1}$	$r_{VIX,t+1}$	σ^2_{t+1}	σ^2_{t+1}
$\Delta_6\sigma_t^2$	-1.21				-0.11	-0.43	-0.23		0.36
	(0.34)				(0.05)	(0.15)	(0.07)		(0.15)
$VIX_t^2 - \sigma_t^2$		4.50							
		(0.62)							
VIX_t^2			-0.14						
			(1.20)						
σ_t^2				-1.36				0.72	
				(0.49)				(0.06)	
σ_{t-6}^2								0.003	
								(0.02)	
N	335	341	341	341	335	264	166	335	335
Adj. R^2	2.8%	7.0%	0.0%	2.0%	3.4%	0.7%	4.6%	51.0%	20.5%

Panel B: Volatility

	Excess Stock Returns				Variance Risk Premium			Volatility	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	$r^e_{M,t+1}$	$r^e_{M,t+1}$	$r^e_{M,t+1}$	$r^e_{M,t+1}$	$VIX_t - \sigma_{t+1}$	$r_{var,t+1}$	$r_{VIX,t+1}$	σ_{t+1}	σ_{t+1}
$\Delta_6\sigma_t$	-0.18				-0.07	-0.084	-0.049		0.34
	(0.11)				(0.04)	(0.035)	(0.014)		(0.12)
$VIX_t - \sigma_t$		0.80							
		(0.20)							
VIX_t			-0.01						
			(0.17)						
σ_t				-0.20				0.73	
				(0.13)				(0.07)	
σ_{t-6}^2								0.04	
								(0.03)	
N	335	341	341	341	335	264	166	335	335
Adj. R^2	1.4%	5.8%	0.0%	1.2%	1.2%	4.2%	0.6%	55.4%	15.1%

Table 2: Stock level analysis. We repeat our results at the stock level. We run three forecasting regressions $y_{i,t+1} = a_i + b\Delta_6\sigma_t^2 + \varepsilon_{i,t+1}$ where $\Delta_6\sigma_{i,t}$ is the 6 month change in realized variance at the stock level for firm i (the realized variance estimates on the right hand size are winsorized at the 95% level, see text for discussion). As dependent variables, y, we use the equity risk premium (stock return over the risk free rate, $r_{i,t+1} - r_{i,t}^f$ labeled ERP), future variance ($\sigma_{i,t+1}^2$), and the variance risk premium (difference between implied variance from option metrics and future realized variance, $VRP_t = IV_{i,t}^2 - \sigma_{i,t+1}^2$ where IV is implied volatility). Data are monthly but realized variance uses daily data with the month. The last three columns repeat the regression using time fixed effects. In our panel regressions standard errors are double clustered by stock and time.

	ERP	Vol	VRP	ERP	Vol	VRP
$\Delta_6\sigma_t^2$	-0.129	0.253^{***}	-0.104^{***}	-0.040	0.188^{***}	-0.070^{***}
	(0.137)	(0.067)	(0.036)	(0.075)	(0.046)	(0.022)
N	536,726	536,726	536,726	536,726	536,726	536,726
Adj R^2	0.001	0.010	0.002	0.159	0.060	0.024
Time FE	N	N	N	Y	Y	Y

Table 3: Survey Expectations. We fit the actual volatility process and the survey expectations to an exponential weighted average on past realized variance. That is, we fit: $\sigma^2_{t,t+k} = a + b\Sigma_{i=1}^{K}\phi^i\sigma^2_{t-i} + \varepsilon_t$ and report the estimated ϕ where we choose K to be 1 year. We then repeat this replacing σ^2_t on the left hand side with the expectation of volatility from the survey. A higher ϕ from the expectations data signifies that expectations rely more on variance farther in the past compared to the optimal forecast for volatility. We use the Graham and Harvey CFO survey (CFO) which is available quarterly and the Shiller survey which is available monthly. Standard errors are below in parentheses.

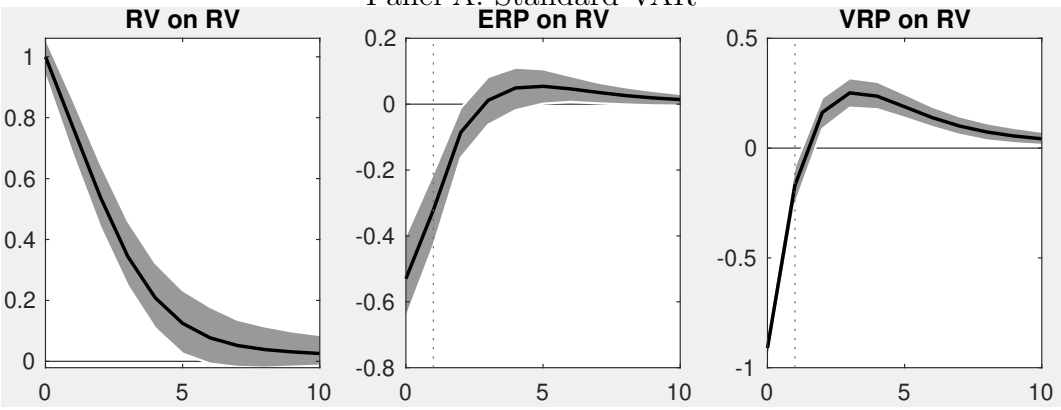
Dependence on Past Variance (ϕ)		
Source	Survey	Future Variance
CFO	0.86***	-0.01
	(0.16)	(0.31)
Shiller	0.73***	0.03
	(0.04)	(0.29)

Table 4: Parameter Estimates from Model (GMM). We estimate the model parameters using two stage GMM. See the text for moments used. Standard errors are in parentheses.

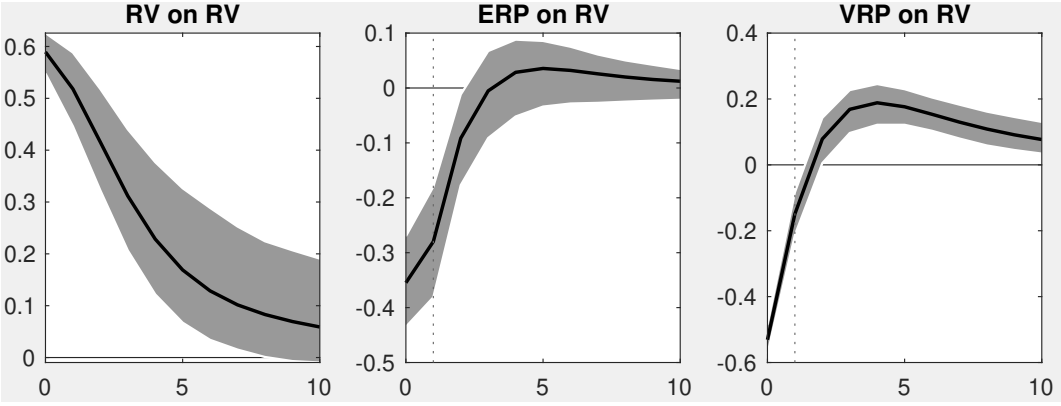
Parameter	Description	Estimate	Standard Error
ϕ	Degree of Extrapolation	0.66	(0.04)
γ	Risk Aversion	5.99	(0.85)
ψ	Elasticity of Intertemporal Substitution	2.87	(1.12)
$\bar{v}$	Unconditional Variance (percent)	0.18	(0.03)
v	Autocorrelation of Variance	0.82	(0.03)
ω	Volatility of Variance Shocks (percent)	0.24	(0.03)

Figure 1: VAR for US data. We run a VAR of realized variance, market excess returns (denoted ERP for equity risk premium), and the variance risk premium (VRP). The variance risk premium is implied variance (from the VIX) minus realized variance. We plot the response of each variable to a shock to realized variance at time 0 but also add a vertical dashed line at time 1 to highlight the predicted, rather than realized, equity and variance risk premiums based on the time 0 RV shock. We include the 6 month lag of each variable in the VAR. The x-axis is in months. Confidence intervals are given in the shaded regions. Panel B weights the observations by the inverse of realized volatility (weighted least squares). Panel C drops the highest 10% of observations based on ex-ante realized variance to include sensitivity to outliers. See text for more detail.

Panel A: Standard VAR



Panel B: Weighted VAR



Panel C: Drop Highest 10% of Variance Realizations

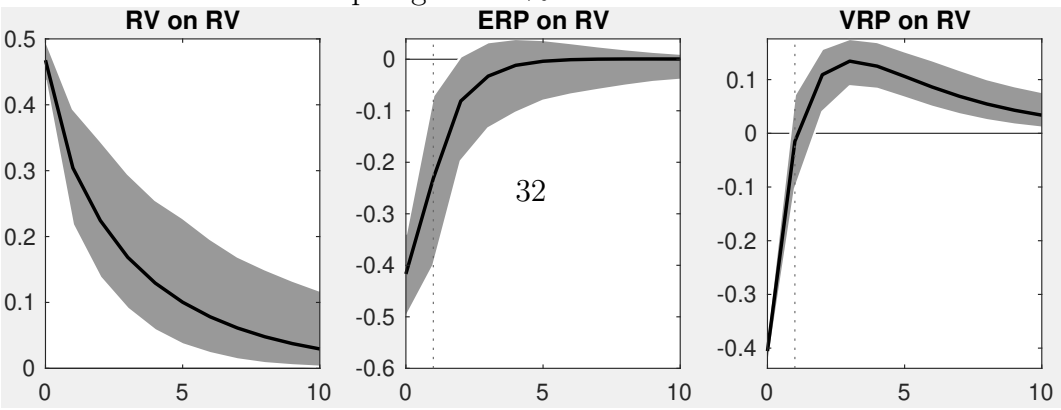


Figure 2: Response to a variance shock in the model. In the top panel we plot the behavior of agents expectations of volatility (blue line) and the true path of expected volatility (dashed red line) in response to an increase in variance in our model. Because agents extrapolate from past volatility they initially underreact and then overreact. The variance risk premium then reflects the difference between agents expectations of volatility minus the rational forecast of volatility, hence it goes negative initially then becomes positive. The bottom panel shows the behavior of stock prices (the price dividend ratio) which also responds slowly in the model, reflecting agents slow moving beliefs. This slow response makes it appear as if equity risk premiums don't initially rise (and potentially even fall) after an increase in volatility but then rise later after the volatility shock has largely subsided. The x-axis is in months.

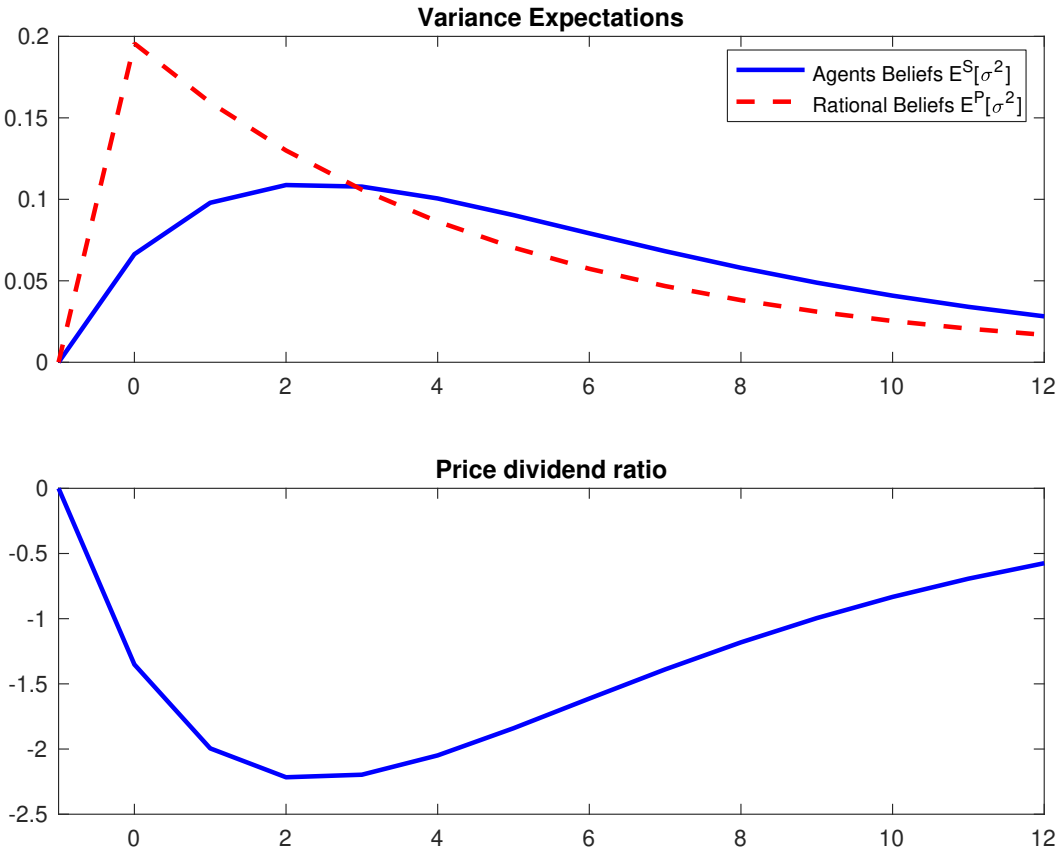


Figure 3: Impulse responses: data vs model. We plot the behavior of expected stock returns and variance risk premiums in the data vs the model at various horizons. The black line shows the impulse response from the data using a VAR of realized variance, excess stock returns, and the variance risk premium. The blue dashed line repeats this using simulated data from the estimated model. The red dot dashed line repeats this exercise in the simulated model data but imposes no extrapolation bias (rational model). The x-axis is in months.

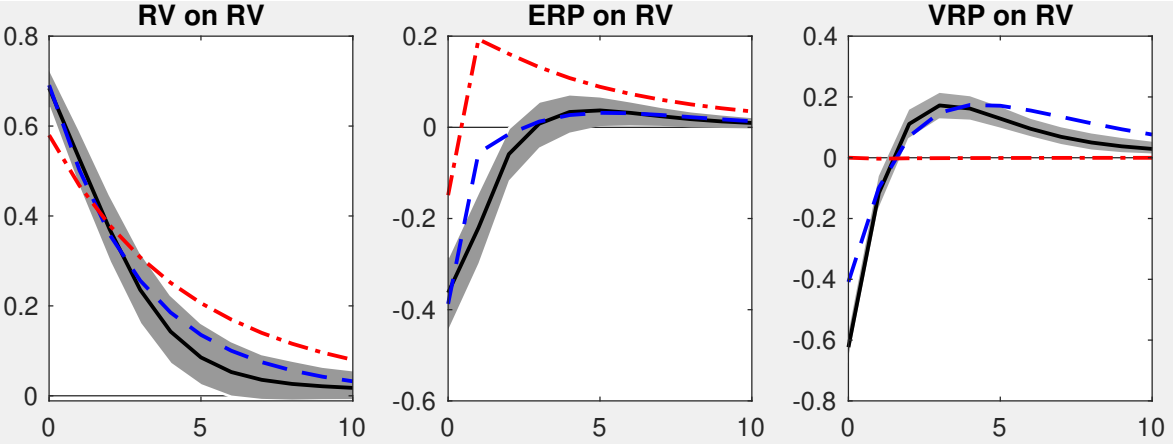


Figure 4: Survey Expectations. We fit the actual volatility process and the survey expectations to an exponential weighted average on past realized volatility. That is, we fit: $\sigma_t^2 = a + b\Sigma_{i=1}^{12}\phi^i\sigma_{t-i}^2 + \varepsilon_t$ and then plot the implied coefficients on each lag from the estimated ϕ. We then repeat this replacing σ_t on the left hand side with the expectation of volatility from the survey. A higher ϕ from the expectations data signifies that expectations rely more on volatility farther in the past compared to the optimal forecast for volatility. We compare the optimal coefficients for forecasting volatility based on past volatility (blue line) vs those implied by the survey data (red dashed line) for the Shiller survey and Graham and Harvey surveys, respectively. Both cases show surveys depend more on volatility in the distant past compared to the optimal.

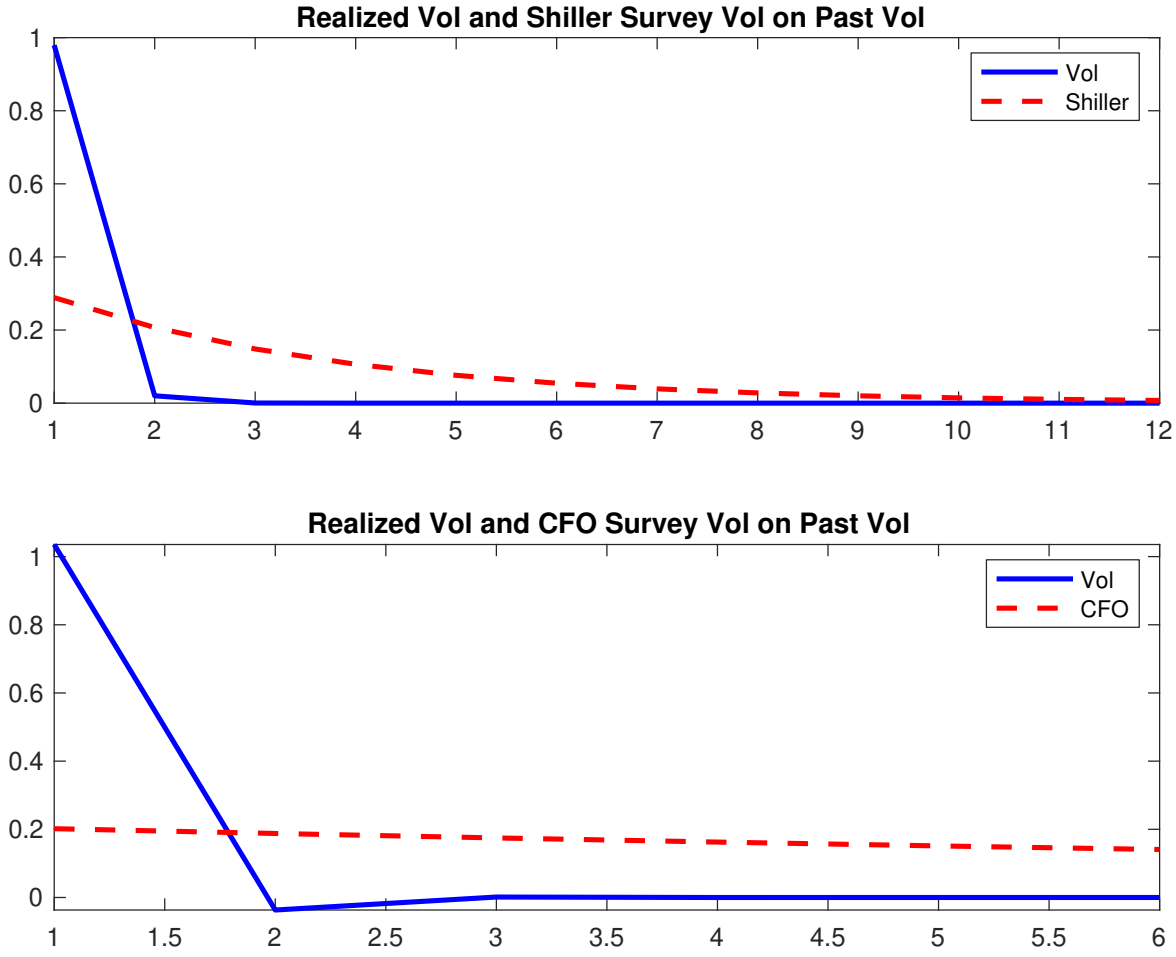
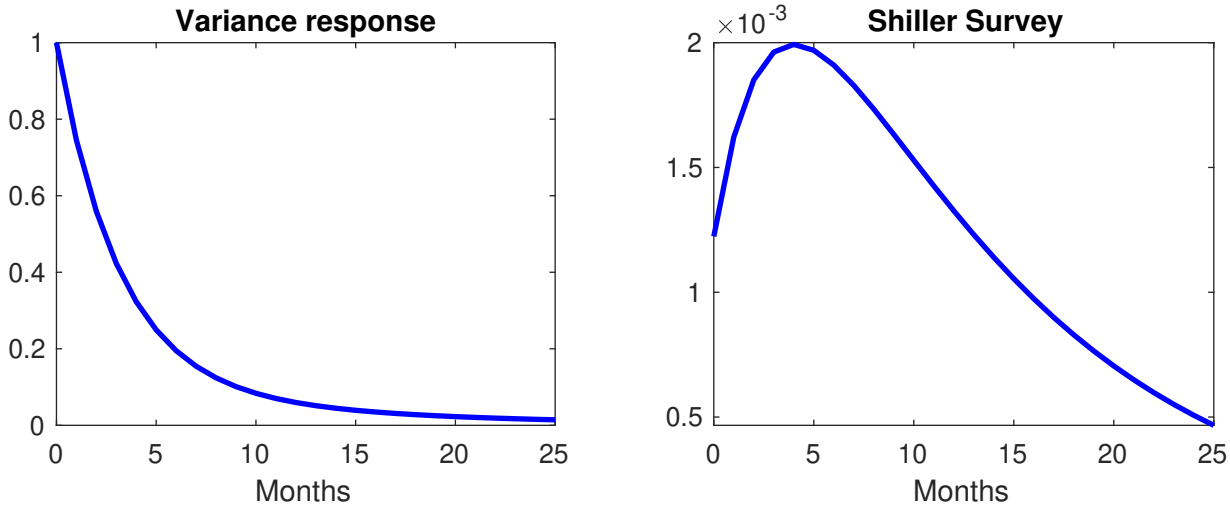
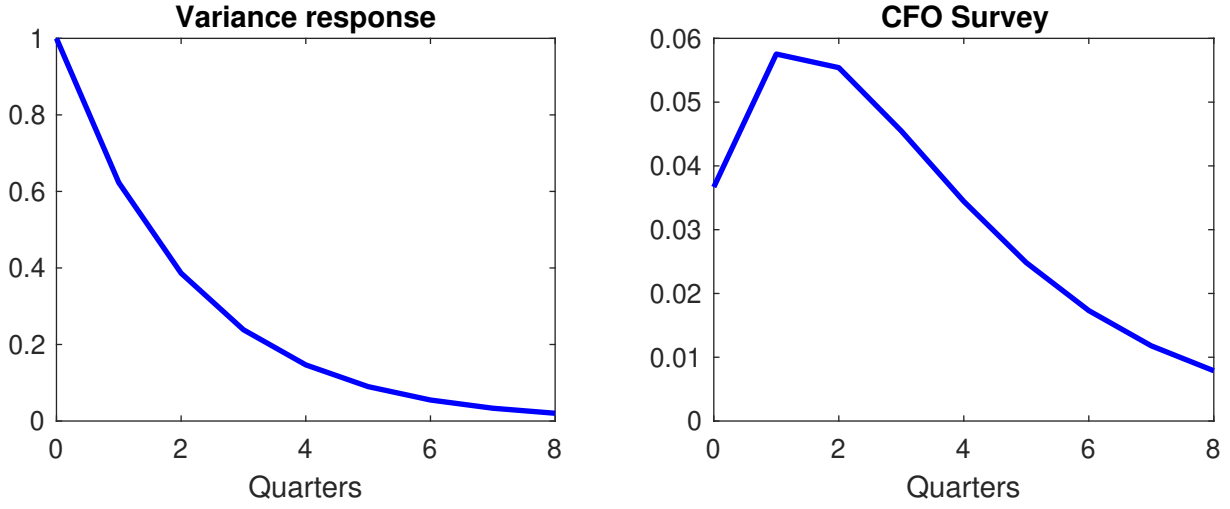


Figure 5: Survey Expectations, VAR. We run a VAR using realized volatility and survey expectations of volatility (in this order) and plot the impulse response to a volatility shock. Expected volatility rises strongly after the shock and then mean reverts fairly quickly. Survey expectations rise slowly, underreact initially and then remain elevated far longer (subsequent overreaction).

Panel A: Shiller Data



Panel B: CFO (Graham Harvey) Data



8. Appendix

Appendix contains additional derivations, tables, and figures.

8.1 Data

The table below details our international data sources including starting time periods for each series.

Table 5: Data Sources.

Country	Index	Volatility	Source	History
USA	SP500	VIX	WRDS	From 1/2/1990
France	CAC 40	VCAC	Bloomberg	Fron 1/3/2000
Canada	sptsx60	VIXC	Montreal Exchange	From 10/2010
UK	FTSE 100	VFTSE	Bloomberg	From 1/4/2000
Germany	DAX	DAX New Volatility (V1XI)	Bloomberg	From 1/2/1992
Japan	Nikkei 225	VXJ	Bloomberg	From 1/5/1998
South Korea	KOSPI	VKOSPI	Bloomberg	From 1/2/2003
Netherlands	AEX	VAEX	Bloomberg	From 1/2000
Switzerland	SMI	V3X	Bloomberg	From 6/28/1999

8.2 Appendix Tables and Figures

Table 6: Change in volatility, equity risk premium, and variance risk premium. We run three forecasting regressions $y_{i,t+1} = a_i + b\Delta_6\sigma_t + \varepsilon_{i,t+1}$ where $\Delta_6\sigma_{i,t}$ is the 6 month change in volatility of the stock market index for country i. As dependent variables, y, we use the equity risk premium (future index return over the risk free rate, $r_{i,t+1} - r^f_{i,t}$ labeled ERP), future volatility ($\sigma_{i,t+1}$), and the volatility risk premium (difference between volatility index and future realized volatility, $VIX_{i,t} - \sigma_{i,t+1}$ labeled VRP). Data are monthly. The first columns use all countries, the last use only US data. In our panel regressions standard errors are clustered by time.

Panel A: Volatility						
	All Countries			US Only		
	(1)	(2)	(3)	(4)	(5)	(6)
	ERP	Vol	VRP	ERP	Vol	VRP
$\Delta_6\sigma_t$	-0.15*	0.27***	-0.06***	-0.25***	0.32***	-0.09***
	(0.09)	(0.08)	(0.02)	(0.08)	(0.05)	(0.02)
N	1,786	1,786	1,786	340	340	340
R-squared	0.01	0.15	0.05	0.03	0.13	0.04
Country	All	All	All	USA	USA	USA
Panel B: Variance						
	All Countries			US Only		
	(1)	(2)	(3)	(4)	(5)	(6)
	ERP	Vol	VRP	ERP	Vol	VRP
$\Delta_6\sigma_t^2$	-0.88***	0.23**	0.04	-1.66***	0.36***	-0.13***
	(0.34)	(0.10)	(0.10)	(0.39)	(0.04)	(0.03)
N	1,786	1,786	1,786	340	340	340
R-squared	0.02	0.10	0.01	0.05	0.19	0.05
Country	All	All	All	USA	USA	USA

Figure 6: Stylized facts for US data. We run regressions of returns, variance risk premiums, and realized variance on lags of realized variance and plot coefficients by horizon. Variance risk premiums are measured either as squared VIX minus realized variance, using the negative for variance swap returns (e.g., selling variance), or using the negative of VIX futures (shorting the VIX). We also plot stock returns on the lagged variance risk premium (squared VIX minus realized variance). The x-axis is in months.

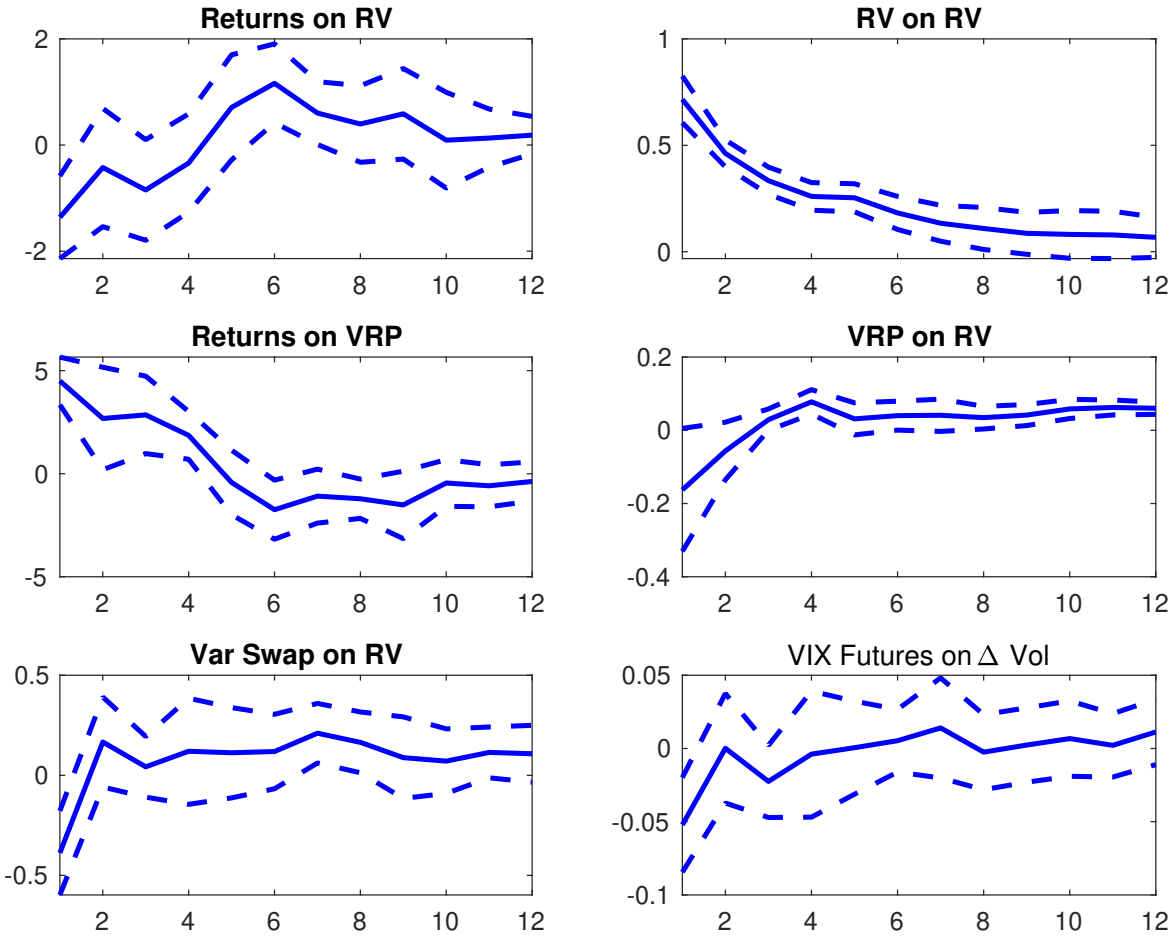


Figure 7: Stylized facts for US data: Post 2010. We replicate our main stylized facts using only US data from 2010 onwards, thus excluding the financial crisis. We run regressions of returns, variance risk premiums, and realized variance on lags of realized variance and plot coefficients by horizon. Variance risk premiums are measured either as squared VIX minus realized variance, using the negative for variance swap returns (e.g., selling variance), or using the negative of VIX futures (shorting the VIX). We also plot stock returns on the lagged variance risk premium (squared VIX minus realized variance). The x-axis is in months.

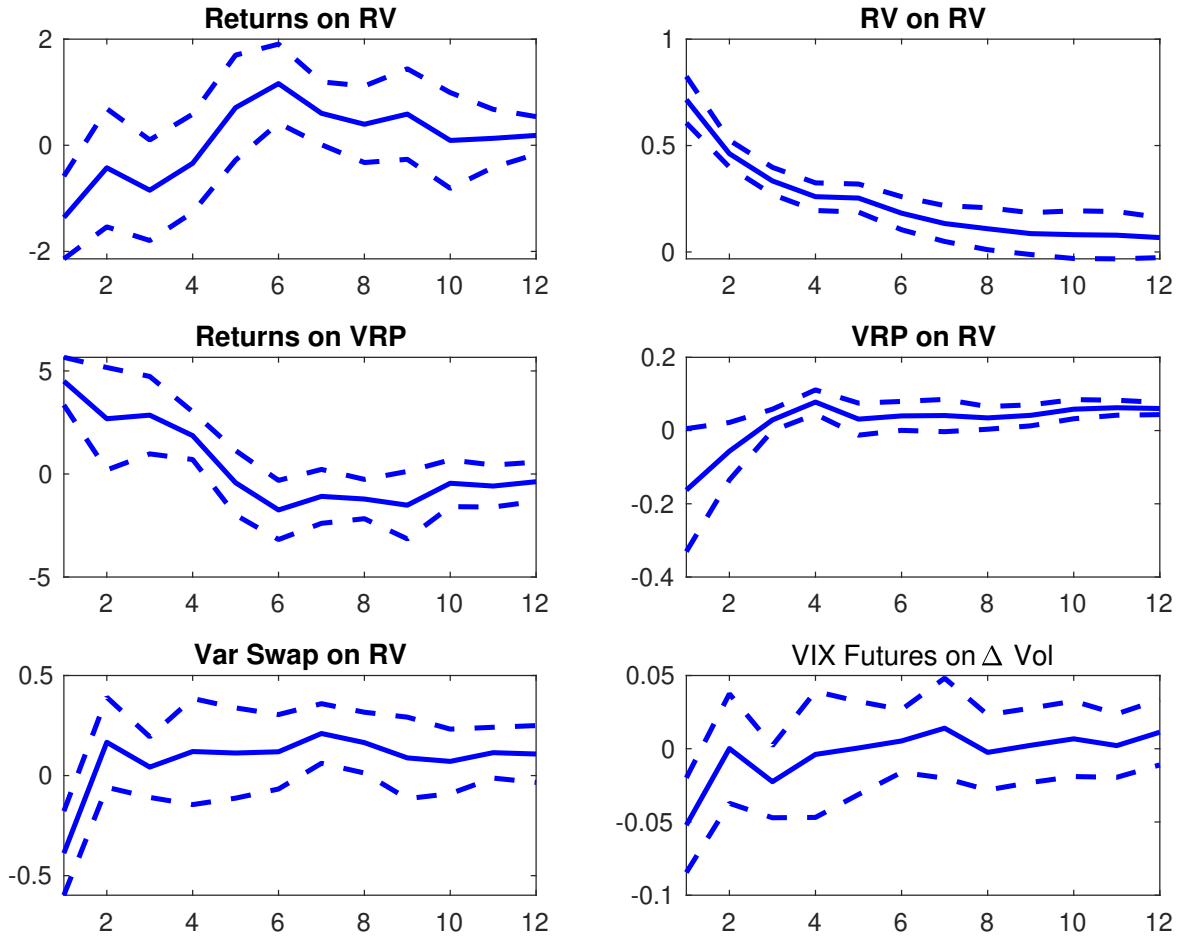


Figure 8: Longer Sample of US Data. We plot regression coefficients of returns on lags of realized variance and of realized variance on lags of realized variance. Our top sample includes all US data from 1926 while the bottom sample includes only post War US data (since 1950). Note the negative coefficient of returns on realized variance is weaker using the longer sample of returns, though even in this sample there is no strong evidence of a positive risk-return tradeoff.